Epidemiology

A Handbook for Medical Students

Prepared for the Community Stream
Faculty of Medicine, University of Colombo

Authors

Prof. Lalini Rajapaksa
Prof. Rohini de A Seneviratne
Prof. Dulitha N Fernando

Edited by

Dr. Carukshi Arambepola
Ms. Kantha Lankatilake

Typesetting by

Dr. Chamara Perera

Recommended Reading

- Hennekens, C.H., and Buring, J.E. *Epidemiology in Medicine*. Boston : Little, Brown & Co., 1987

- Beaglehole, R., Bonita, R., Kjellström, T. *Basic Epidemiology*. Geneva: World Health Organization, 1993

- Hulley, S.B., and Cummings, S.R. *Designing Clinical Research: An Epidemiological Approach*. Baltimore: William & Wilkins, 1988.

- Leon, Gordis. *Epidemiology* (3rd edition), 2004.

1. Introduction to Epidemiology

1.1 What is 'epidemiology'?

controling factors

Epidemiology is the study of the **distribution** and **determinants** of **disease frequency in human populations.**

This definition is based on 2 fundamental assumptions:

- Human disease does not occur at random.
- Human disease has causal ~~casual~~ and preventive factors that can be identified through systematic investigations of different populations in different places at different times.

The application of epidemiology is therefore to control health problems in a defined population. This emphasizes that epidemiologists are concerned not only with death, illness and disability but also with positive health states and ways of improving health.

1.2 What are the aims of epidemiology?

- To describe the distribution and magnitude of health and disease problems in human populations
- To identify the causes/correlates (related factors) of disease
- To provide data essential for planning, implementation and evaluation of services for prevention, control and treatment of diseases and to prioritize those services

1.3 Measurements in epidemiology

In describing the distribution and magnitude of health and disease problems in a population, certain measurements are required. They are,

- measurements of mortality – no of deaths
- measurements of morbidity – disease appears in population
- measurements of disability – dissabled populatyon
- measurements of natality no of birth.
- measurements of the characteristics or attributes of a disease
- measurements of medical needs, health care facilities and its utilization

- measurements of environmental factors related to disease
- measurements of demographic variables

Count
Proportion
Percentage
Rate Ratio

1.4 Types of measures used in epidemiology

Measurements in epidemiology are made using different measures, which quantify the occurrence of disease or health related events. The most basic measure of disease frequency is a **COUNT** of affected individuals.

Example

During an epidemic of dengue haemorrhagic fever in 2007, 150 new cases were reported from area A and 85 cases from area B.

Does this information help you to decide where it would be safer to live in?
The answer to this question is NO.

We cannot draw any conclusions as we do not know the size of the population in the two areas, from which these cases were reported.

A count alone has limited value in describing the "problem" in a community.
The count needs to be related to the size of the source population in which the cases or events occurred.

Let us now see the information presented in Table 1.

Table 1: No. of dengue haemorrhagic fever cases and population in areas A and B

Location	No. of new cases	Population
Area A	150	37,500
Area B	85	7,100

Although many cases were reported from area A, it is noted that the population of this area is five times that of area B.

In area A, 150 out of 37,500 got dengue i.e. one out of every 250 persons.

In area B, 85 out of 7,100 got dengue i.e. one out of every 84 persons.

Here, we have taken the **PROPORTION** of dengue cases in the total population of the area. This proportion can also be expressed as a **PERCENTAGE.** For example,

Area A:
The percentage of dengue cases among its total population = 150 / 37,500 x 100
= 0.4%

Area B:
The percentage of dengue cases among its total population = 85 / 7,100 x 100
= 1.2%

It appears that area A is better to live in than area B.

However, the picture would be different if the 85 cases in area B occurred during a period of one year, compared to 150 cases occurring in area A over a period of 4 months.

Now, calculate the percentage of dengue cases in areas A & B over one year and comment on your findings.

Time factor, constant

Percentage of dengue cases in areas A over one year $= \frac{150 \times 3}{37500} \times 100$

$= 1.2\%$

Percentage of dengue cases in area B over one year $= \frac{85}{7100} \times 100$

$= 1.19\%$

Dengue cases are approximately equal in 2 areas.

What does all this mean?

When measuring disease or a health-related event in a community, it is necessary to use a measure that will take into account the number of cases, the population in which such cases occur and the time period during which the cases occurred.

RATE is such a measure of disease frequency.

Proportion
Ratio –

- A rate measures the occurrence of a given event in a defined population during a given period of time. Rates are therefore calculated by expressing the number of events (numerator) as a fraction of the population (denominator), in which the event occurred.

$$\text{Rate} = \frac{\text{No. events occurring in a given population during a given time period} \quad (n)}{\text{No. of subjects in that population in which the event occurred} \quad (d)} \times K$$

The rate comprises of the following:

1. • *numerator,*
2. • *denominator,*
3. • *time specification during which the event occurred and* (duration)
4. • *a constant K, which is usually a multiple of 10.*

You have learnt about rates when you studied demography.

Please read up the module on demography. You may recall that we followed the same principle i.e. the numerator being the number of events and the denominator being the population in which the event can occur and the time period.

RATIO is another measure of disease frequency. It expresses a relation in size between two 'counts' obtained by simply dividing one quantity by another.

For example, male: female ratio in a defined population is given by the number of males to number of females. **X: Y or X / Y**

You should note here that the numerator is not a component of the denominator. Now think of other ratios that you know of.

Children : adults
Doctor : patient
Dependency ratio.
Urban population : rural population
Doctor : population ~~rati~~

What are the measures of disease frequency that you learnt so far and how do these differ from each other? Rate, Proportion, Percentage, count, Ratio. 2 unrelated things.

Exercise 1

Q1. The following information is given for district A for the year 2007.

Total population	1.7 million
Geographical extent	652 sq. kilometers
Male population	51 %
Urban population	1.2 million

A survey carried out in the district revealed that 777,240 men were able to read and write.

Calculate the following:

1. Population density
2. Sex ratio — per hundred male how many females.
3. Proportion of urban population
4. Literacy rate for the male population

1. Population density = $\frac{1.7 \times 10^6}{652}$ = 2607 per km² in district A.

2. Sex ratio = Male : Female 51 : 49   Sex ratio = $\frac{49}{51} \times 100 =$

3. Proportion of urban population = $\frac{1.2}{1.7}$ = 0.705

4. Male population = $1.7 \times 10^6 \times \frac{51}{100} = 8.67 \times 10^5$

Literacy rate for the male population = $\frac{777240}{8.67 \times 10^5} \times 100$

= 89.6

89 male out of 100 in area A in 2007

Q2. This area had 273,000 households, of which 29,000 households had no toilet facilities. The water supply to this area was through roadside taps and from wells. A total of 14,450 deaths were reported from the area, of which 1,068 were infant deaths. Of these infant deaths, 75% occurred during the first month of life. Twenty percent of total deaths in the area were due to diarrhoea.

A total of 42,700 live births were recorded in this area. This area being a proclaimed area where the Registrar of Births is a medical person, still births were also registered. A total of 700 still births were registered. There were 5 maternal deaths reported during this year.

2.1 Calculate the following rates using the above data

1. Crude death rate = $\frac{\text{Total No of deaths}}{\text{Total population}} \times 1000$
2. Crude birth rate
3. Infant mortality rate = $\frac{\text{No of infant deaths}}{\text{Total live births.}} \times 1000$
4. Maternal mortality rate
5. Neonatal mortality rate = $\frac{\text{No of neonatal deaths}}{\text{No of live births}} \times 1000$
6. Cause specific mortality rate for diarrhea = $\frac{\text{No of deaths from diarrhoea}}{\text{Total population.}} \times 100$
7. Still birth rate

of Live Births / population × 1000

aternal deaths / births. × 100,000

ll births × 1000 / births (live + still)

2.2 List the main observations you could make regarding the health status of district A.

1. Crude death rate = $\frac{14450}{1.7 \times 10^6} \times 10^3$ = 8.5 for 1000 population in 2007

2. Crude birth rate = $(42700 / 1.7 \times 10^6) \times 1000$ = 25.1 for 1000 population in 2007

3. Infant mortality rate = $(1068 / 42700) \times 10^3$ = 25.01 infant death for 1000 live births.

4. Maternal mortality rate = $(5 / 42700) \times 10^5$ = 11.7 maternal deaths for 10^5 live birth

5. No. of neonatal deaths = $1068 \times \frac{75}{100}$ = 801

Neonatal mortality rate = $(801 / 42700) \times 10^3$

= 18.75

6. Cause specific mortality ~~Deaths due to diarrhoea~~ = $\frac{14450}{5}$ = 2890

Cause specific mortality rate for diarrhoea = $(2890 / 1.7 \times 10^6) \times 10^5$ = 170

7. Still birth rate = $\frac{700}{43400} \times 10^3$ = 16.1 still birth for per 1000 births.

Have to compare with national rates. (Annual health bulletin)

Q3. There are 12 Medical Officers of Health; 226 Public Health Midwives; and 26 Public Health Nursing Sisters in this area. A total of 1,165 Medical Officers and 3.165 Nurses work in the 25 hospitals in the area. During year 2007, 10,000 beds were available and 500,000 patients were treated as inward-patients in these hospitals. OPD attendance was 3,940,000.

Calculate the following:

1. Availability of health personnel on a population basis
2. Hospitals per population
3. Beds per population
4. OPD visits on a population basis

1. • Medical officers of health per 100 000 population = $\frac{12}{1.7\times10^6}\times10^5$ =

• Public health midwives per 100 000 population = $\frac{226}{1.7\times10^6}\times10^5$ =

• Public health nursing sisters per 100 000 population = $\frac{26}{1.7\times10^6}\times10^5$ =

• Medical officers per 100 000 population = $\frac{1165}{1.7\times10^6}\times10^5$ = 68

• Nurses per 100 000 population = $\frac{3165}{1.7\times10^6}\times10^5$ = 186

2. Hospitals for 100 000 population = $\frac{25}{1.7\times10^6}\times10^5$ = 1.4

3. Beds per 1000 population = $\frac{10\,000}{1.7\times10^6}\times10^3$ = 5.8

4. OPD visits on 1000 a population = $\frac{3\,940\,000}{1.7\times10^6}\times10^3$

= 2 317

2. Prevalence and Incidence

The measures of disease frequency most commonly used in epidemiology fall into two broad categories:

① Disease frequencies -
- **Prevalence**
- **Incidence**

Measuring prevalence and incidence depends on counting 'cases' in a defined population within a defined time period.

The first question we must ask ourselves when we attempt to count disease is, "Who is a 'case'?" i.e. the **case definition.**

This may sound very simple but let us look at the following example.

Cholera can exist in the population in the following manner:

	→		→		→	
Stools demonstrate *Cholera vibrio* No signs and symptoms	→	Few loose stools Person does not take much notice	→	Mild watery diarrhoea May/ may not seek treatment	→	Typical 'rice water' stools Needs hospitalization

Which of these would we count as a 'case of cholera'?

Any stage can be taken as a case of cholera after defining ~~the~~ what we are taking.

It is very clear that before we start quantifying disease, we need a clear definition of a 'case' to identify 'disease' / 'healthy status" in a person. The definition should be **clear** and **unambiguous.**

> In epidemiology, it is essential that the definition of 'case' be clearly stated.
>
> It should be easy to use and easy to measure in a standard manner under a wide variety of circumstances by different people.

The next question we much ask ourselves when we attempt to count disease is, "Do we have a correct estimate of the population in which a 'case' could occur?"

Ideally, this source population should include only persons who are susceptible to that illness. For example, in the case of food poisoning, only those who consumed the infected food will form the population at risk. The section of the population that is susceptible is called the **population at risk**. Sometimes, information on the population at risk is not always available and in such a situation, the total population is used as an approximation.

In a visual examination of patients for identifying the prevalence of cataract, whom do you consider as 'cases' and 'population at risk'?

✓ Cases - People diagnosed as having cataract by the visual examinatio
Population at risk - × Total population in a area during a
particular period
✓ Those above 50 years. Those have undergone cataract
Surgery & those who are blind should be excluded.

2.1 Prevalence

In quantifying disease in a defined population, we can count all individuals who have the disease at a given point in time. This number in the defined population is called the **prevalence.**

Prevalence is a measure of disease burden in a community / population as it gives an idea of the disease status in a defined population at a given point in time.

$$\textbf{Prevalence (P)} = \frac{\text{No. of existing cases of a disease or condition at a specified point in time}}{\text{Population-at-risk in the defined population at the same point in time}} \times K$$

K = multiple of 10

Since we are referring here to a specified point in time, this is also called **point prevalence.** In this instance, prevalence does not as such involve a time period. Therefore, by strict definition, prevalence is a proportion and not a rate. However, since the point can also refer to a specific point in calendar time such as, per week, per year etc., it is also called a **prevalence rate.**

2.2 Incidence

Incidence of a disease is the **new** cases of a disease that occur **during a specified period of time** in a defined population.

Incidence rate is defined as the number of new cases occurring in a defined population-at-risk during a given period of time.

$$\textbf{Incidence Rate (IR)} = \frac{\text{No. of new cases of a specified disease during a given period of time}}{\text{Population-at-risk during that time period}} \times K$$

K = multiple of 10

It can also refer to new spells or episodes of illness that occur during the given period of time. For example, if a person suffers from a common cold twice during the year, there would be 2 spells of sickness in that year.

$$\textbf{Incidence Rate (Spells)} = \frac{\text{No. of new spells or episodes of a specified disease during a given period of time}}{\text{Population-at-risk during that time period}} \times K$$

K = multiple of 10

Incidence measures the rate at which cases occur in a population. It is not influenced by the duration of the disease.

If a count of all cases i.e. old cases plus new cases that occur over a period of time (i.e the total number of persons who are known to have had the disease at any time during a specified period of time) is taken, a period prevalence can be calculated. The denominator used for this calculation is the population-at-risk midway through the defined period of time.

$$\textbf{Period Prevalence} = \frac{\begin{array}{c}\text{No. of existing cases of a specified disease} \\ \text{at the beginning of a given period} \\ + \\ \text{No. new cases diagnosed during the same period}\end{array}}{\text{Estimated at-risk population at 'mid' time interval}} \times K$$

K = multiple of 10

This measure combines both point prevalence (status at a single point in time) and incidence (risk of developing disease over a period of time) in a single parameter.

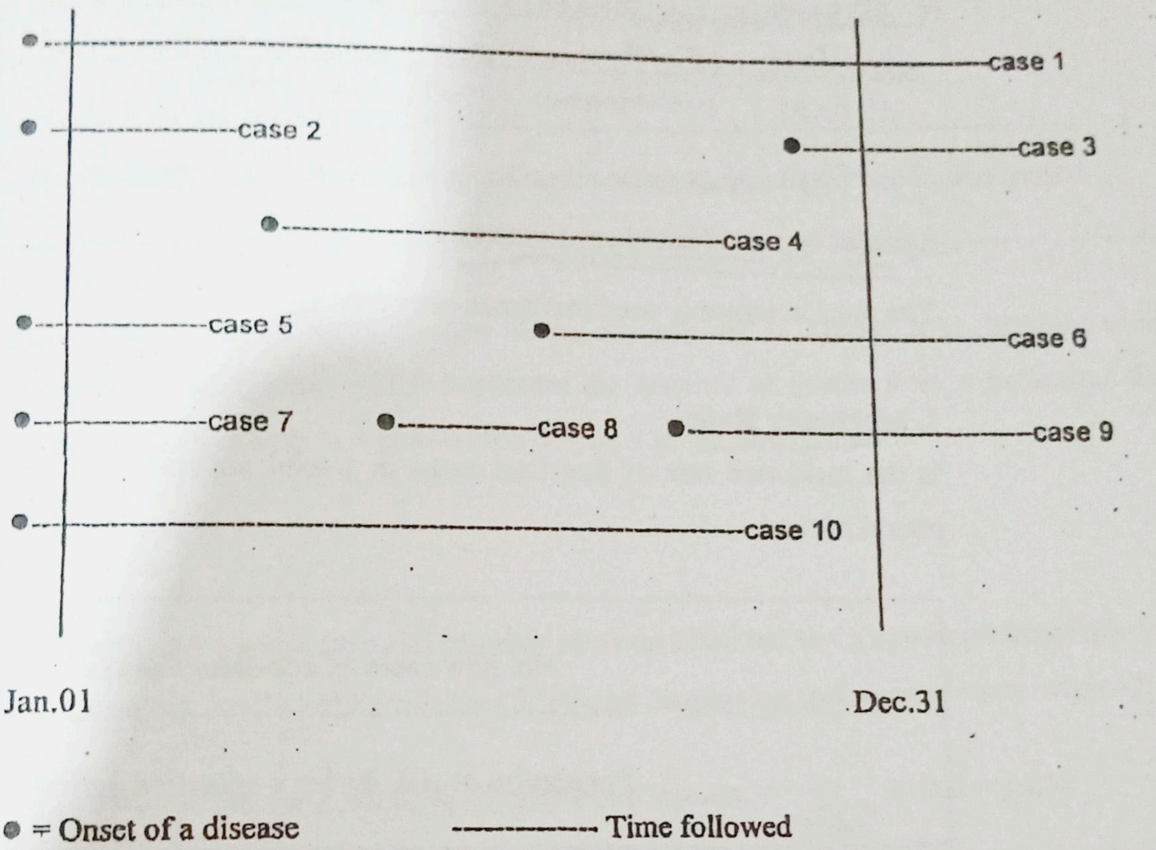
This method is useful and convenient when it is particularly difficult to determine when a disease can be considered present. Such as in the diagnosis of mental illness.

In period prevalence, what are the two disease frequencies that are combined in a single parameter.

Point prevalence & incidence

Example

Figure 1 illustrates 10 persons at the beginning and ending of a disease during a given time period



Point prevalence on Jan.01 = 1,2,5,7,10

Point prevalence on Dec.31 = 1,3,6,9

Incidence (Jan 1-Dec 31) = 3,4,6,8,9

Period prevalence (Jan 01 – Dec 31) = (1,2,5,7,10) + (3,4,6,8,9)

Example

In a study of oral contraceptive (OC) use and bacteriuria, a total of 2390 women aged 16-49 years were identified as free from bacteriuria. Of these, 482 were OC users at the time of the first survey in 1985. At the second survey in 1988, 27 of the OC users had developed bacteriuria.

What is the incidence rate of bacteriuria among OC users?

IR = 27 cases of bacteriuria among 482 OC users

= 27 / 482 or 5.6 percent bacteriuria among OC users within 3 years

It is a risk rather than a rate.

The time specification is important since 5.6 percent in 6 months or 1 year is quite different to 5.6 percent in three years. IR assumes that the entire population-at-risk at the beginning of the time period (in this instance, 482 OC users) has been followed up for the three years.

More about incidence . . .

The incidence that we have been discussing so far is also called a **cumulative incidence (CI)**. Our counts of the new disease episodes are those that have been accumulated over a defined period of time. Thus, cumulative incidence provides an estimate of the likelihood or risk of an individual developing the disease during the specified period of time.

Here, we assume that the total group is available throughout the study period, hence we are able to identify disease occurrence for the whole group within the total duration. However, this may not always be so. For example, if we have a group of 1000 persons in our study group, who will be studied during a period of one full year, some of them (e.g. 250) are available in the area only for 6 months, that is for 0.5 years, whereas the other 750 are available for 1 year.

We have one of two options:

1. To leave out those who were not available for 'one full year' and calculate incidence based only on the 750 persons or
2. Find a way to use the data related to the total group. If so, how do we do this?

We can calculate the **person-time** during which each member in the group was followed up. i.e. 250 for 0.5 years and 750 for one year

Therefore, the total person-time of observations is = (250 x 0.5) + (750 x 1)

As you can see the person-time of observation can be used to account for the varying time periods of follow up. Using this, we could now calculate another type of incidence called **'incidence density'**.

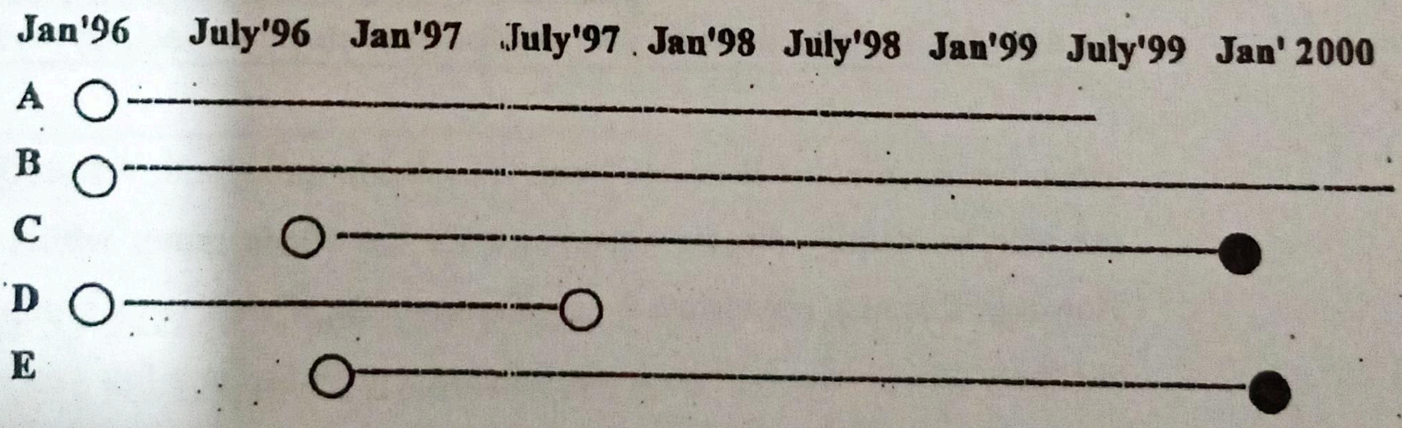
$$\text{Incidence Density (ID)} = \frac{\text{No. of new cases of a disease during a given period}}{\text{Total person-time of observations}} \times K$$

K = multiple of 10

Let's take a look at the example given in Figure 2.

Example

Figure 2 shows that different subjects A - E have been followed up for varying periods of time.



A-E subjects considered ○

Time followed ———

On set of the disease ●

For example, subject A was followed up for 3 years; subject B for 4 years etc.

Total number of person-years = A - 3; B - 4; C - 3; D – 1; E - 3 = 14 person years

Number of cases = 2

Can you think of the factors that may influence the prevalence of a disease in a defined population? Age, Environmental factors, Co-morbidity, Hot Changes in hormones, Gender, Lack of Genetic factors.

✓ Incidence of the disease (if rare, Prevalence will be common prevalance will be high.)

✓ Duration of illness (if highly fatal or quick recovery, Pr be low. if chronic disease, prevalence will be high illness, prevalence will be low.)

Prevalence = Incidence × Duration of illness

Special types of incidence rates:

The most commonly used incidence rates are:

1. Morbidity Rate:

Is the incidence rate of non-fatal cases in a total population at risk during a specified period of time.

$$\text{Morbidity rate} = \frac{\text{No. new cases of non-fatal disease}}{\text{Population-at-risk during a specified period of time}} \times K$$

K = multiple of 10

2. Mortality Rate:

Is the incidence rate of fatal cases (deaths) in a total population during a specified period of time.

$$\text{Mortality rate} = \frac{\text{No. of deaths from a disease}}{\text{Total population}} \times K$$

K = multiple of 10

3. Case Fatality Rate (CFR):

CFR measures the number of deaths that occur from a specific illness in a group of patients suffering from that illness in a given period of time. It represents the proportion of fatal cases among affected individuals and is often expressed as a percentage.

$$\text{Case Fatality Rate (CFR)} = \frac{\text{No. deaths from a disease in a given period of time}}{\text{No. diagnosed cases of that disease in that period}} \times K$$

Case fatality rate is an indication of the **severity** of an illness. It is different to ~~cause~~ case specific mortality rate, which expresses the number of deaths from a particular disease among all individuals in a population during a given time period.

Example

In 1995 in Sri Lanka, the total number of cases admitted to Government hospitals with poisoning by medicinal agents was 2590 and deaths due to the same cause were 47.

Case fatality rate for poisoning by medicinal agents, $= \frac{47}{2590} \times 100 = 1.8\%$.

i.e. 18 deaths per 1000 cases of poisoning by medicinal agents in Sri Lanka in 1995

It should be noted that the deaths that make up the numerator in CFR do not necessarily represent the cases that make up the denominator.

Can you suggest an example to support this statement? When the person died from when a disease which was diagnosed before that period eg: DM

---

4. Attack rate (AR):

AR is a measure of occurrence of new cases of a disease in a specific population during a short period of time, as in an outbreak of a disease, often due to a specific exposure.

$$\text{Attack Rate (AR)} = \frac{\text{No. of new cases of a disease exposed to a specific exposure during a short period of time}}{\text{Population-at-risk in that limited period of observation}} \times K$$

- This measure is useful in investigating an acute epidemic. E.g. Food poisoning among a group of pilgrims during Poson season in Mihinthale area

In an outbreak of food poisoning, it is not enough to calculate only the attack rate for the illness but also to identify the food item/ s responsible. This is done by examining the number of individuals who ate different types of foods and determining the frequency of occurrence of illness among those who ate and those who did not eat each food item.

The difference between the attack rates for those who were exposed and not exposed to a particular food item provides important clues in the investigation of aetiology of an acute outbreak

Exercise 1

Q1. An investigation of an outbreak of gastro-intestinal illness in Laswego county in New York revealed that 37 out of 60 people who had attended a dinner became ill within a few hours. Given below is an epidemiological analysis of this outbreak.

Table 1. Characteristics of persons in Laswego county during an outbreak of acute-gastroenteritis

Age (yrs)	Sex	Time of eating food	Date of onset	mutton	chicken	rice	fish	dhal	coffee	water	tea	vanila	chocolate
11	M	unknown	well										Y
52	F	8.00 p.m.	19 0.30 am	Y	Y	Y			Y			Y	
65	M	6.30 p.m.	19 0.30 am	Y	Y	Y	Y		Y			Y	Y
69	F	6.30 p.m.	19 0.30 am	Y	Y		Y	Y	Y		Y		
13	F	unknown	well										
63	F	7.30 p.m.	18 10.30 pm	Y	Y		Y	Y	Y		Y		
70	M	7.30 p.m.	18 10.30 pm	Y	Y		Y	Y	Y	Y		Y	
40	F	7.30 p.m.	19 2.00 am							Y			Y
15	F	10.00p.m.	19 1.00 am								Y		Y
33	F	7.00 p.m.	18 11.00 pm	Y	Y	Y			Y			Y	Y
65	M	unknown	well	Y	Y	Y							
38	F	unknown	well	Y	Y	Y			Y		Y		Y
62	F	unknown	well	Y	Y		Y	Y		Y			
10	M	7,30 p.m.	19 2,00 am									Y	Y
25	M	unknown	well	Y	Y	Y	Y	Y	Y	Y	Y	Y	
32	F	unknown	19 10.30 am	Y	Y				Y		Y	Y	Y
62	F	unknown	19 12.30 am									Y	
36	M	unknown	18 10.15pm	Y	Y		Y					Y	
11	M	unknown	well	Y	Y		Y			Y			Y
33	F	unknown	18 10.00 pm	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
35	F	unknown	well									Y	Y
48	F	unknown	19 2.00 a.m	Y	Y	Y	Y	Y	Y		Y	Y	Y
25	M	unknown	18 11.00 p.m	Y	Y					Y	Y	Y	Y
11	F	unknown	well										Y
74	M	unknown	18 10.30 p.m	Y	Y	Y	Y	Y	Y		Y	Y	
12	F	10.00 p.m	19 1.00 a.m								Y	Y	Y

Age (yrs)	Sex	Time of eating food	Date of onset	mutton	chicken	rice	fish	dhal	coffee	water	tea	vanilla
44	F	7.30 p.m	19 2.30 a.m	✓						Y	Y	
3	M	unknown	18 9.45 p.m	✓	Y	Y				Y	Y	
53	F	7.30 p.m	18 11.30 p.m	Y	Y	Y	Y	Y	Y	Y	Y	
13	F	10.00 p.m	19 1.00 a.m	✓							Y	
77	M	unknown	18 11.00 p.m	Y	Y	Y	Y	Y		Y	Y	
64	M	unknown	well									
65	F	unknown	well	Y	Y	Y	Y	Y	Y	Y	Y	
59	F	unknown	18 9.45 p.m	✓	Y	Y	Y			Y	Y	
15	F	10.00 p.m	19 1.00 p.m	✓							Y	Y
62	M	unknown	well	Y	Y		Y		Y	Y	Y	Y
37	F	unknown	18 11.00 p.m	Y	Y	Y		Y	Y		Y	Y
17	M	10.00 p.m	well								Y	Y
35	M	unknown	well	Y	Y	Y		Y	Y		Y	Y
15	M	10.00 p.m	19 9.00 p.m	✓							Y	Y
50	F	unknown	well									Y
40	M	unknown	well	Y	Y				Y	Y	Y	
35	F	unknown	well	Y	Y	Y			Y	Y		
35	F	unknown	18 9.15 p.m	Y	Y	Y	Y		Y		✓	Y
36	M	unknown	well	Y		Y	Y		Y			
57	F	unknown	18 11.30p,m	Y	Y		Y	Y	Y		Y	Y
16	F	10.00 p.m	19 1.00 a.m	✓							Y	
68	M	unknown	18 9.30 p.m	Y		Y	Y		Y		✓	Y
54	F	unknown	well	Y	Y	Y			Y		Y	
77	M	unknown	19 2.30 a.m	✓							✓	Y
72	F	unknown	19 2.00 a.m	Y	Y		Y	Y	Y		Y	Y
58	M	unknown	18 9.30 p.m	Y	Y	Y			Y		✓	Y
20	M	10.00 p.m	well								Y	Y
17	M	unknown	well	Y	Y	Y				Y		Y
62	F	unknown	19 12.30 a.m	Y	Y					Y	✓	Y
20	F	7.00 p.m	19 1.00 a.m	✓							✓	Y
52	F	unknown	18 10.30 p.m	Y	Y	Y	Y			Y	✓	Y
9	F	unknown	well								Y	
50	M	unknown	well	Y	Y	Y	Y	Y	Y	Y	Y	
8	M	11.00 p.m	19 3.00 a.m	✓							✓	Y

The investigation revealed that among the 40 persons who had eaten vanilla ice-cream, 31 became ill compared with only 6 out of 20 who did not eat the ice-cream. The attack rates for those who ate and did not eat vanilla ice cream are as follows:

AR (Vanilla ice-cream +) = 31/40 x 100 = 77.5%

AR (Vanilla ice-cream -) = 6/20 x 100 = 30%

1.1 Calculate the total attack rate.

1.2. Calculate the attack rate by sex.

1.3 Calculate the attack rates for those who ate and those who do not eat each food item.

1.1 Total attack rate = $\frac{37}{60} \times 100$ = 6.1%

61 cases per 1000 people at risk in that limited period of observation

~~Not eaten~~

1.2. Attack rate (by sex) = $\frac{13}{262} \times 100$ = 50% AR
males

Attack rates (female) = $\frac{6}{34} \times 100$ = 20.5%

Eaten | Not Eaten

1.3. Attack rate (for mutton) = $\frac{22}{36} \times 100$ = 61.1% AR = $\frac{15}{24} \times 100$ = 62.5%

AR (for chicken) = $\frac{23}{36} \times 100$ = 63.8% AR = $\frac{14}{24} \times 100$ = 58.3

AR (for rice) = $\frac{15}{25} \times 100$ = 60% AR = $\frac{22}{35} \times 100$ = 62.85

AR (for fish) = $\frac{16}{23} \times 100$ = 69.5% AR = $\frac{21}{37} \times 100$ = 56.75%

AR (for dhal) = $\frac{11}{16} \times 100$ = 68.75% AR $\frac{26}{44} \times 100$ = 59.09%

AR (for coffee) = $\frac{17}{27} \times 100$ = 62.96% AR = $\frac{20}{33} \times 100$ = 60.6%

AR (for water) = $\frac{11}{20} \times 100$ = 55% AR = $\frac{26}{40} \times 100$ = 65%

AR (for tea) = $\frac{21}{32} \times 100$ = 65.62% AR = $\frac{16}{28} \times 100$ = 57.14%

AR (for vanilla) = $\frac{31}{40} \times 100$ = 77.5% AR $\frac{6}{20} \times 100$ = 30%

AR (for chocolate) = $\frac{19}{35} \times 100$ = 54.28% AR = $\frac{18}{25} \times 100$ = 72%

Q2. The number of cases and deaths due to shigellosis in Sri Lanka from 1984 -1995 based on admissions to government hospitals (excluding Northern and Eastern provinces) is given in Table 2. Answer the questions given below.

Table 2. Number of cases and deaths due to shigellosis in Sri Lanka from 1984-1995 based on admissions to government hospitals (excluding Northern and Eastern provinces)

Year	Morbidity		Mortality		Case Fatality Rate (%)
	No.	Rate per 100,000	No.	Rate per 100,000	
1984	15,896	101.9	147	0.94	0.92
1985	12,667	81.2	126	0.81	0.99
1986	18,256	117.0	127	0.81	0.69
1987	23,806	152.6	150	0.96	0.63
1988	28,284	181.3	243	1.56	0.66
1989	30,070	193.4	277	1.78	0.92
1990	28,284	181.2	180	1.15	0.64
1991	31,334	200.8	172	1.10	0.55
1992	42,693	273.7	185	1.19	0.43
1993	31,272	200.4	91	0.58	0.29
1994	27721	177.6	76	0.49	0.27

Population of Sri Lanka excluding those of Northern & Eastern Provinces 15,599,600 (156×10^5)

2.1 Calculate the morbidity, mortality and case fatality rates for each year and complete Table 2.

2.2 Comment on the trends of shigellosis.

Morbidity ↑ - Inadequate health care

mortality ↓ - ORS.

✓ Morbidity rates have increased over time → ORS programme preventing deaths

Moratality rates have decreased over time → Inadequate health education of on hand washing, inadequate water & facilitutes for safe drinking. sanitary facilities increasing the morbidity.

CFR has decreased over time due to above two.

Inadequate facilities for sanitary facilities

Q3. Figures 3 and 4 represent the follow up of 7 subjects with exposure to a risk factor and the time taken for the development of a disease.

Figure 3. Outcome of a follow up of 7 exposed person-population A

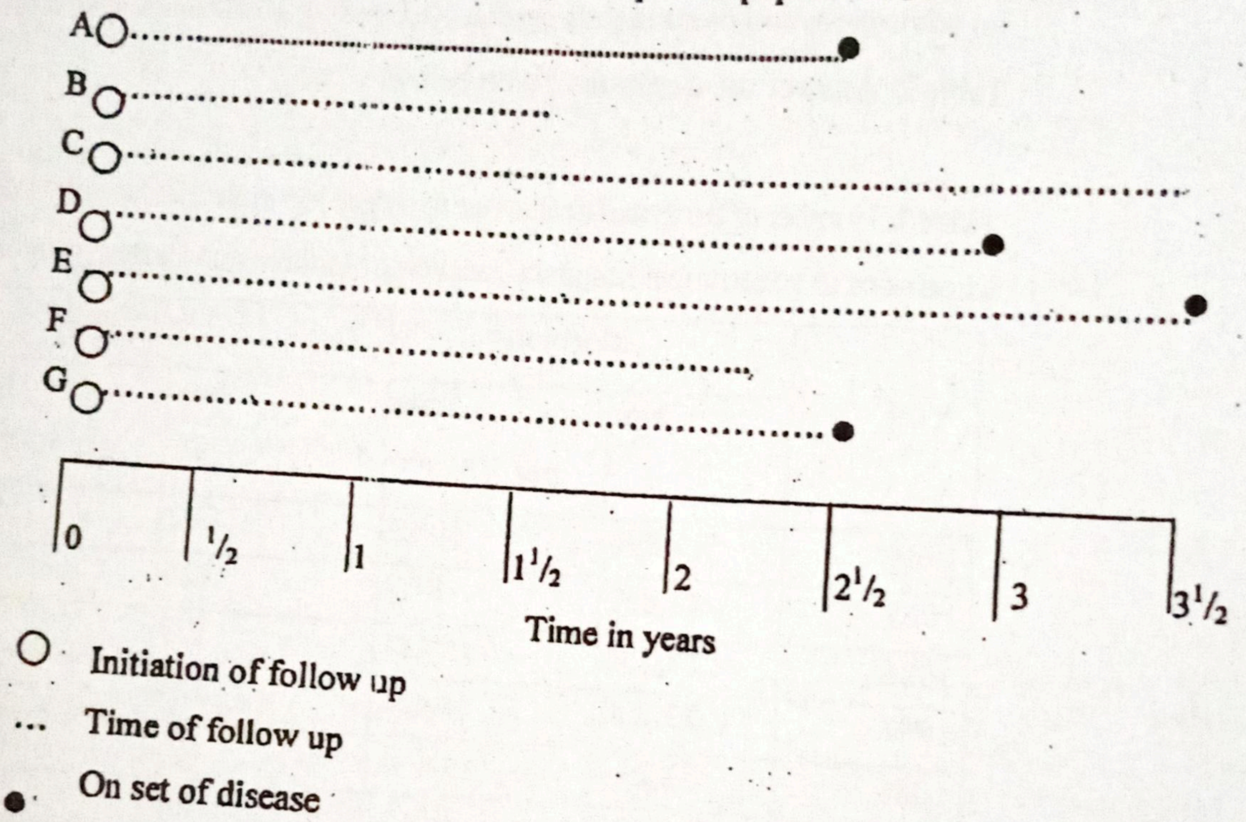
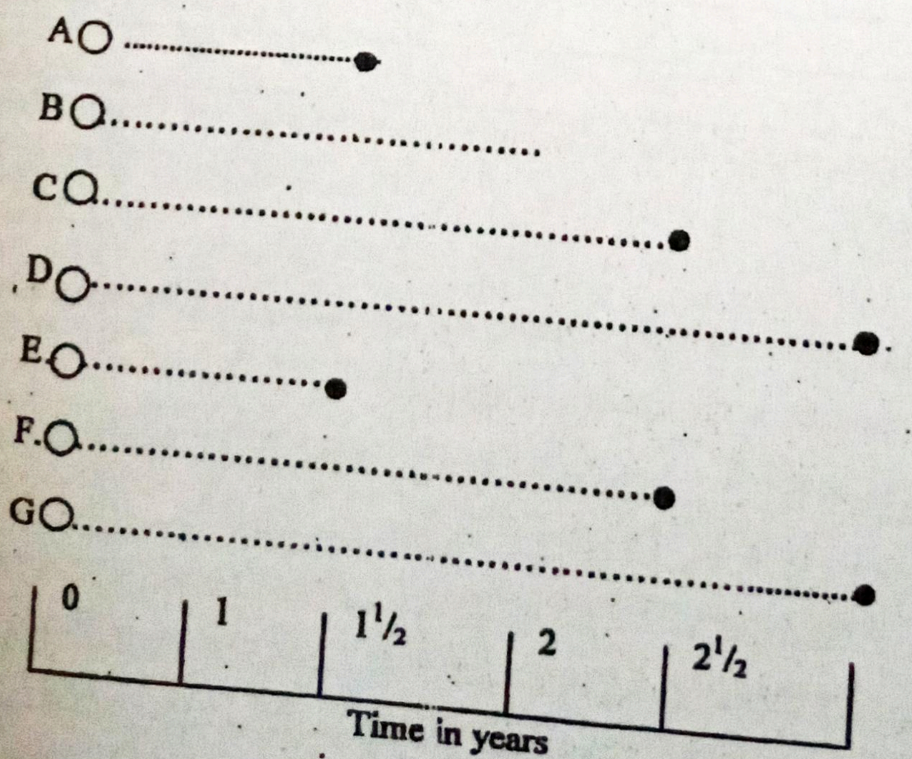


Figure 4. Outcome of a follow-up of 7 exposed persons in population B



3.1 How many new cases occurred in populations A and B?

A - 4
B 6

3.2 Calculate the person-time of exposure in years for:

Population A....$2\frac{1}{2} + 1\frac{1}{2} + 3\frac{1}{2} + 3 + 3\frac{1}{2} + 2\frac{1}{2} + 2\frac{1}{4} =$ 18.5 + 0.25

Population B....$1\frac{1}{2} + 1\frac{1}{2} + 2 + 2\frac{1}{2} + 6 + 2\frac{1}{2} + 1\frac{1}{2}$ = 16

3.3 Calculate the incidence density (Incidence rate) for:

Population A....$\frac{4}{18.5} \times 100 =$ 21.62%

Population B....$\frac{6}{16} \times 100 =$ 37.5%

3.4 Comment on the findings.

Risk of developing disease is higher in population B.

Q.4. In a 5 year follow up study of post menopausal hormonal use and coronary heart disease (CHD), 90 cases with CHD were diagnosed among 32,317 post menopausal women on Hormonal Replacement Therapy (HRT) during a total of 105,786.2 person years of follow up.

4.1 Calculate the incidence rate and incidence density among the participants in this study.

Incidence rate = $\frac{90}{32317} \times 100$ =

Incidence density = $\frac{\text{No of new cases}}{\text{Total person time}} \times 1000$

$= \frac{90}{105,786.2} \times 1000$ (PYS)

= 8.5

Pl. note: It is assumed that risk of CHD in the group remains constant over time. This may not be true since individuals have varying degrees of risk. The ways of overcoming this problem will be dealt with later.

3. Epidemiological study designs

Epidemiological study designs are of two types. They are,

- observational studies
- experimental studies

3.1 Observational studies

Observational studies are those in which data are gathered simply by observing events as they happen, without the investigator playing an active role in what is taking place.

> Observational studies allow nature to take its course; the investigator **observes and measures** but does not intervene.

There are two types of observational studies:

3.1.1. Descriptive studies

3.1.2. Analytical studies

3.1.1 Descriptive studies

> Descriptive studies focus on **describing** the occurrence of a disease or health related event in a population in terms of person, place and time.

Often a descriptive study is the first step in an epidemiological investigation.

3.1.2 Analytical studies

> Analytical studies focus on the **determinants** of a disease or health related event in order to establish whether a particular exposure causes or prevents it.

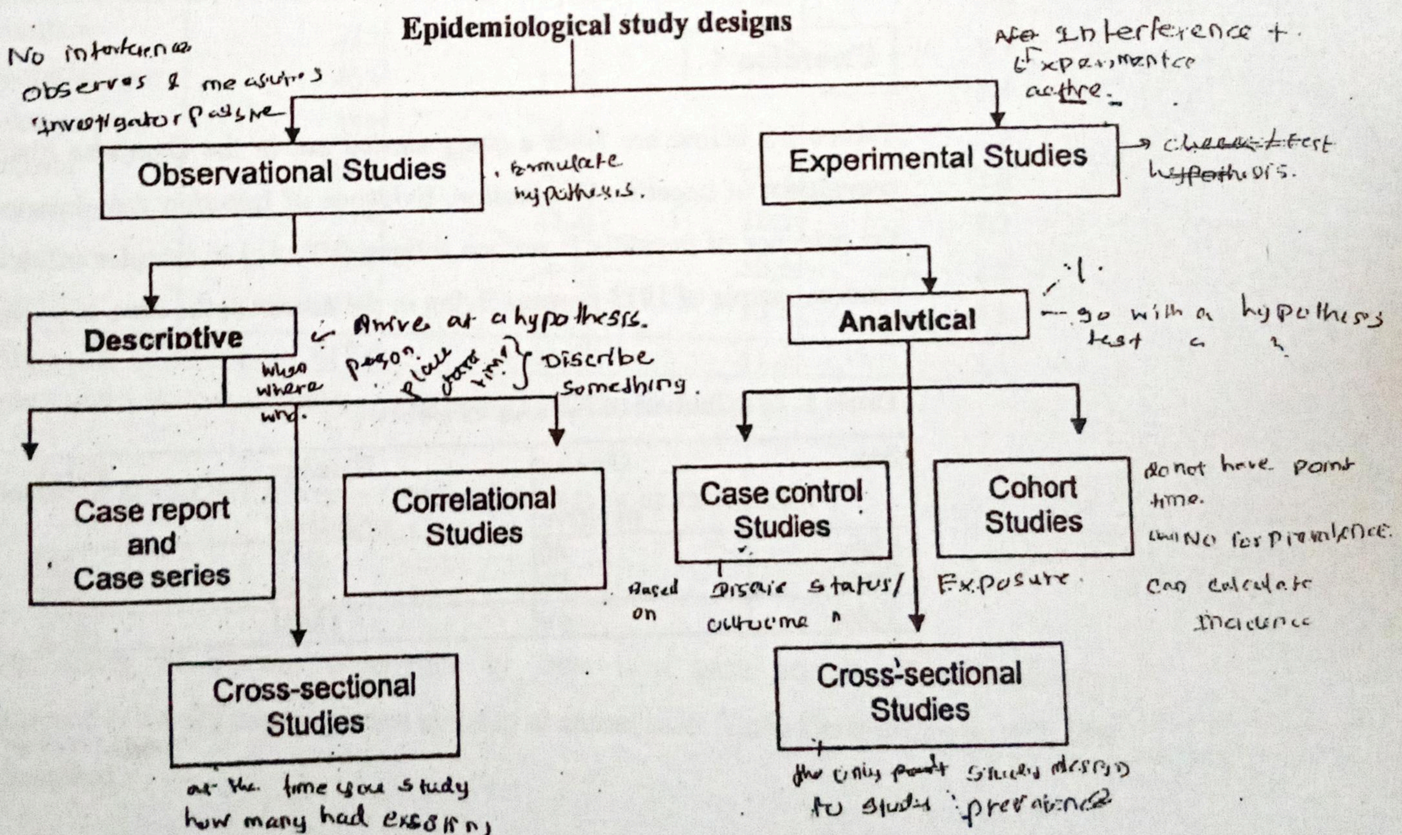
3.2 Experimental studies

> In experimental studies, the investigator **'intervenes' (makes a change) in one or more variables in a group and 'does not intervene' in another group.** This 'change' introduced by the investigator may vary.

Examples: introducing a new drug, introducing a training programme. Here, the study attempts to find out whether the intervention had an influence on a defined outcome. Therefore, experimental studies are also called **intervention studies.**

Summary

Given below is a summary of what we have discussed up to now.



at the time you study how many had exposure ~~for~~ disease.?

can check family hx.)

How many variables you can measure? any amount you want.

We can't do not know which preceded which. (disadvantage)

3.1 Observational studies

3.1.1 Descriptive studies

As the name suggests, these studies describe the amount and the distribution of disease patterns in populations. To describe the occurrence of disease fully, some questions have to be answered:

- **who** *(person)* get the disease?
- **when** *(time)* do cases occur ?
- **where** *(place)* do cases occur ?

In other words, the description has to tell us the **person, place** and **time** in respect of a disease occurrence.

Person: Characteristics of a person such as age, sex, religion, marital status, socio-economical factors, etc. can furnish different types of clues about the pattern and possible etiology of a disease.

Let us work on Exercise 1 that is given below to illustrate the above statement.

Exercise 1

Tables 1-5 below are from a study carried out in the Gampaha district to describe the prevalence of hepatitis B infection. Evidence of infection was determined by testing for the presence of hepatitis B surface antigen (HBsAg) in samples of blood obtained from a random sample of 1913 persons living in the district at the time of survey.

Table 1. Distribution of HBsAg by gender

Sex	HbsAg positive	HBsAg negative	Total
Male	40	947	987
Female	08	918	926
Total	48	1865	1913

Table 2. Distribution of HBsAg status by age

Age in years	HbsAg positive	HBsAg negative	Total
0-4	10	110	120
5-9	12	233	245
10-14	3	192	195
15-19	4	149	153
20-29	7	353	360
30-39	6	344	352
40-49	3	232	235
50 and over	3	250	253
Total	48	1865	1913

Table 3. Distribution of HBsAg status by marital status

Marital status	HbsAg positive	HBsAg negative	Total
Married	18	982	1002
Unmarried	30	883	911
Total	48	1865	1913

Table 4. Distribution of HBsAg status by family characteristics

Family characteristics	HbsAg positive	HBsAg negative	Total
Number of persons 5 or below	15	915	930
Number of persons more than 5	33	950	983
Number of children 0-2	16	505	521
Number of children 3 or more	32	44	481

Table 5. Distribution of HBsAg status by socio-economic status

Socio-economical status	HbsAg positive	HBsAg negative	Total
Upper	04	336	340
Middle	31	1286	1317
Low	13	243	256
Total	48	1865	1913

1.1 What is the measure of disease frequency that can be calculated from the findings of this survey?

~~Rate~~ Period Prevalance

1.2 Calculate the above mentioned measure of disease frequency.

1.3 Describe the pattern of hepatitis B infection in the Gampaha district (Use extra paper).

Place: Frequency of disease can be related to a place of occurrence in terms of areas described by natural boundaries such as rivers, mountains or desserts or by political boundaries such as districts, Grama-Niladhari divisions. Characteristics of the physical and biological environment of an area may cause certain diseases to be more common in that area..

Comparison of disease frequency in relation to place can also be made between countries or between regions within a single country. For example, mortality from colon cancer is much lower in Japan than in USA (Ref. Doll R and Peto R. *The causes of cancer.* New York: Oxford University Press 1981).

Think of possible reasons for this difference.

Exposure to risk factors might be higher in USA than Japan.
Management of patients better in Japan than USA
Preventive measures may be effective in Japan than USA.

Exercise 2

Table 6 is reproduced from a report on the national nutritional survey carried out in 1988/1989 in Sri Lanka..

Table 6. Nutritional status by districts in SL during 1988/89

District	Sample size	% Chronic malnutrition	% Acute malnutrition	% Concurrent malnutrition
Colombo	292	13.7	9.9	2.1
Gampaha	324	8.6	5.6	0.7
Kalutara	327	19.3	8.0	1.2
Kandy	448	38.4	9.2	4.6
Matale	341	19.9	17.6	5.4
Nuwara-Eliya	348	22.7	11.5	2.9
Galle	393	18 1	10.7	3.6
Matara	363	13.8	11.8	2.8
Kurunegala	293	11.9	8.9	0.8
Puttalam	315	16.8	9.5	0.4
Anuradahapura	453	17.0	13.9	4.4
Polonnaruwa	383	18.0	15.1	2.9
Badulla	1060	30.0	10.2	3.8
Moneragala	314	25.2	18.5	7.2
Ratnapura	268	22.8	10,8	2.2
Kegalle	223	23.3	13.0	3.9
Total	6172	21.6	11.4	3.1

Source: Report on the National Nutritional Status Survey in SL, 1988/89

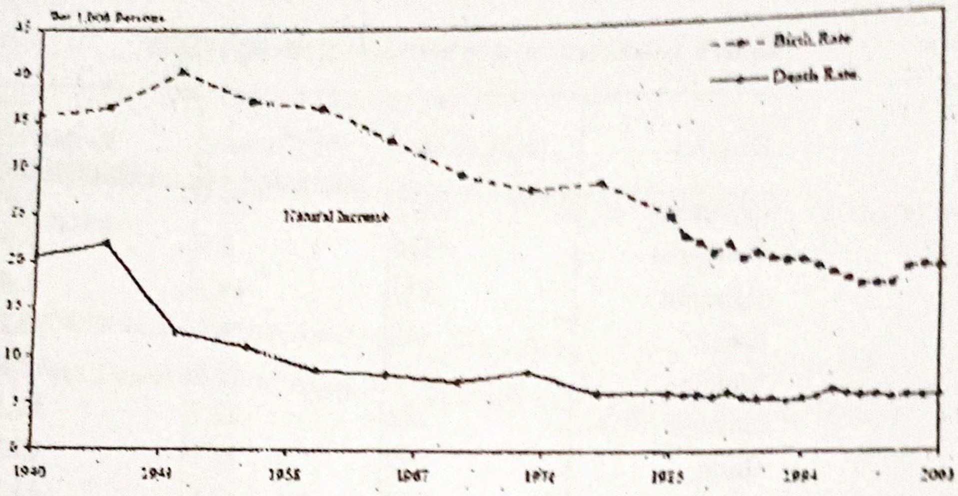
Describe the district differences in nutritional status (Use extra paper).

Time: Study of disease occurrence by **time** is a basic aspect of epidemiology. Occurrence is usually expressed on a monthly or annual basis. Three kinds of change with time are described.

3 kinds of change with time

The first consists of long term variation and is called **secular trends.** This implies changes in the occurrence of disease (i.e. a progressive increase or decrease) over a long period of time, generally over several years or decades. The crude birth and death rates for Sri Lanka from 1945 to 2003 (as shown in Figure 1) shows the five yearly births/deaths rates depicting a secular trend.

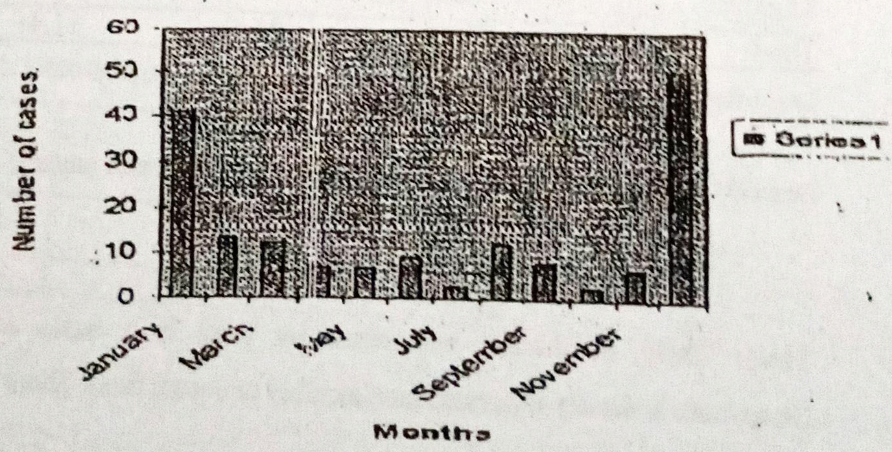
Figure 1. Secular trends in crude birth and death rates of Sri Lanka (1945-2003)



Source: Registrar General's Department

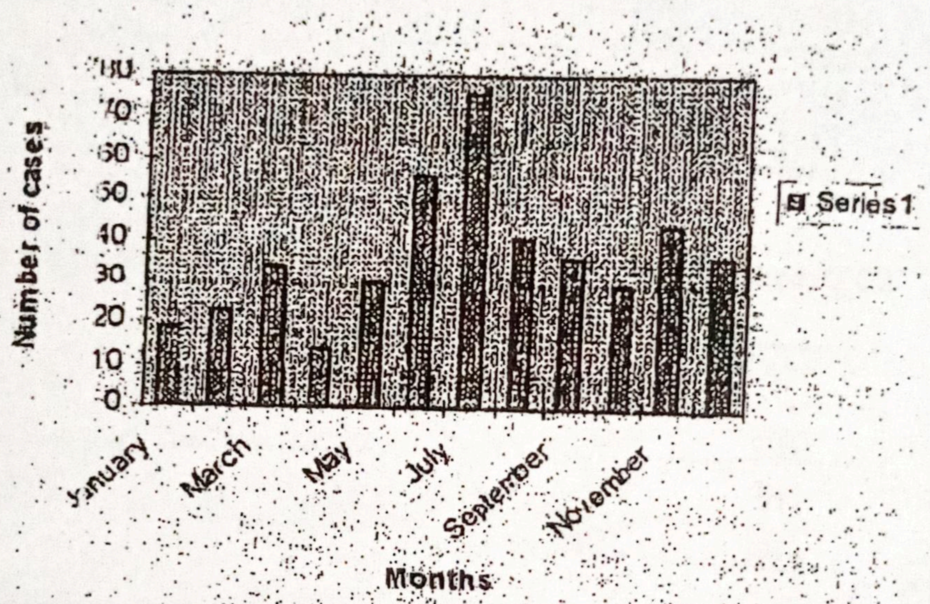
Within these secular-trends, periodic fluctuations may be seen. These could be on an annual or on a seasonal basis and are called **cyclical changes**. A classic example is given in Figure 2.

Figure 2. Distribution of cases of Japanese Encephalitis by months during 1995



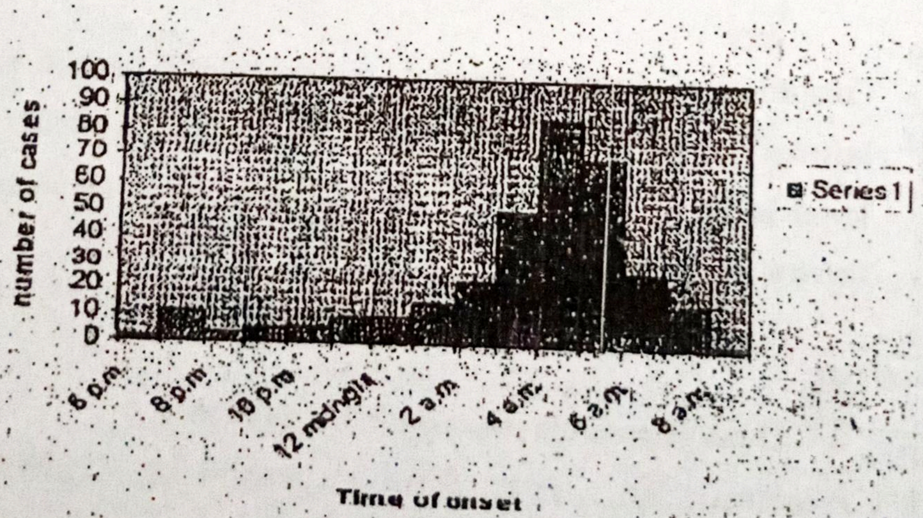
Short term fluctuations, as seen in an epidemic is another type of variation of di frequencies with time. This pattern is called an **epidemic curve.** Two examples are gi below in Figures 3 and 4.

Figure 3. Distribution of Dengue Haemorrhagic Fever by months during 1995



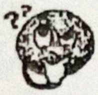
Exercise 3

Figure 4. Epidemic curve for an outbreak of Food Poisoning



Summary

Discussed above were some of the commonly used ways in which descriptive epidemiology summarizes its basic data on health, disease and death in a systematic fashion.



Can you now think of different uses of descriptive studies?

To [illegible] outbreaks or epidemics in [illegible] & [illegible] in a particular period.

To [illegible] a better living [illegible] to live.

[illegible] diseases [illegible]

To identify time in which people are more vulnerable to get diseases

To identify people who are more susceptible [illegible] diseases in health [illegible]

To identify disease frequencies

Descriptive studies use data from many diverse sources. Can you suggest some common sources of data for descriptive studies?

Pry - Collecting data by questionnaires from people.

IIry - from registers & reports without reaching people.

Several approaches are used to obtain data for descriptive studies. They are,

a. Case reports and case series

b. Correlation studies

c. Cross sectional surveys

a. Case reports and case series

A **case report** is the most basic type of descriptive studies consisting of a carefully detailed report of a *single patient*. These describe the experience of a single patient. New or unusual collection of individual case reports gives rise to a **case series**, which describes the characteristics of a number of patients with a given disease. These help to identify unusual clinical presentations of a disease and may lead to formulation of a hypothesis about a possible cause.

Routine surveillance programs often use accumulating case reports to suggest the emergence of new diseases or epidemics. Let's look at the following classic example:

Five young previously healthy homosexual men were admitted to hospital with a diagnosis of penumocystis carinii pneumonia in three hospitals in Los Angeles, during a six month period from 1980-1981.

This clustering of cases was striking because previous to this, infection with *Pneumocystis carinii* had been reported exclusively among older men and women whose immune systems were suppressed. This suggested that these young men were suffering from a previously unknown disease which causes immunosuppression. This was later called Acquired Immune Deficiency Syndrome (AIDS). The fact that all cases were homosexual men also raised the hypothesis that some aspects of sexual behaviour could be related to the risk of acquiring this disease.

What limitations will a researcher face when interpreting the results of a case report/series?

- There is no comparison group to analyse
- Populations are not studied, only individuals.
- can't prove a hypothesis,
- Can't study large numbers.

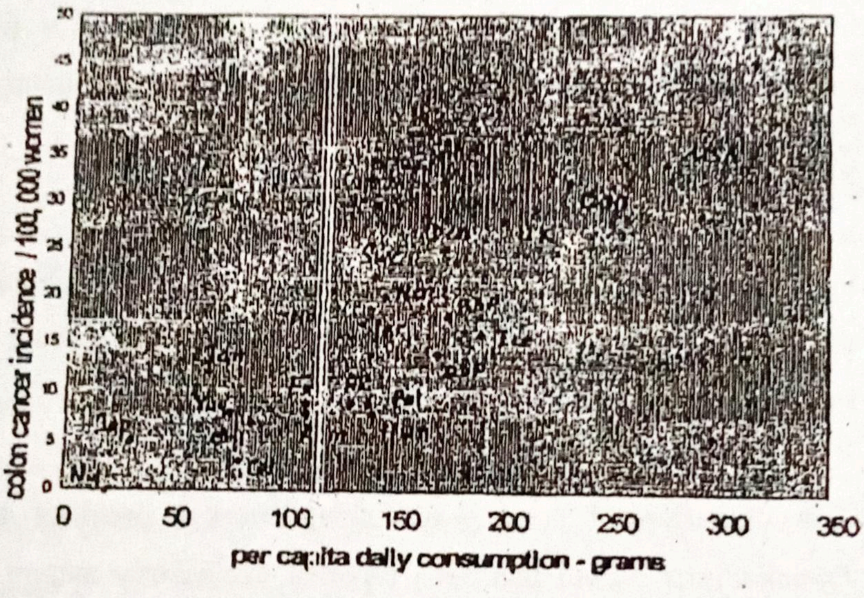
Not getting sample, but sample
size exposure & outcome

b. Correlational studies

Unlike in case series, data from population groups are used to compare disease frequencies. Comparisons are made using the same population at different periods of time or different groups of populations during the same period of time.

Fig 5 refers to the correlation between per capita meat consumption and colon cancer among women in various countries.

Figure 5. Correlation between per capita meat consumption and colon cancer among women in different countries



What associations can you derive from this data?

Correlational studies are used as the first step in investigating a possible relationship between 'exposure' and 'disease' and formulating hypotheses. Other advantages are that they can be done quickly and inexpensively, often using already available data.

The main limitation of correlational studies is its inability to link an exposure to the occurrence of a disease in the same person as the data refers to population groups and therefore not used to test hypotheses. Another limitation of correlation studies is that, there may be other factors between countries that are associated with the exposure, which may account for the differences in disease frequency.

What other factors could be associated with meat consumption and its association with colon cancer? Understand these limitations.

Fat content

Genetic predisposition

Radiation

c. Cross sectional surveys

The third type of descriptive epidemiological design is cross sectional survey. These studies are carried out at a single examination in a **cross section** of the population. Such studies provide information about frequency and characteristics of a disease by providing a 'snapshot' of the experience of a population at a specified period or point in time.

Such information can be of great value to public health administrators in assessing the health status and health care needs of a population. Please refer to Exercises 1 and 2 in Chapter 3 for such cross sectional surveys.

In conducting a cross sectional study, one must:

- have a clear objective
- define the study population
- give due consideration to using a sampling method
- ensure adequate response rates
- identify ' methods' of data collection
- carry out appropriate analysis

i. What is the measure of disease frequency that could be calculated in a cross sectional study?

Prevalence

ii. What kind of cases (incident or prevalent) would a cross sectional survey identify?

Prevalent cases

* Cross sectional study is the method to determine the point or period prevalence of a disease or attribute.

Exercise 4

Q1. It has been observed in cross sectional studies, that individuals with cancer have significantly lower levels of beta carotene than healthy individuals of the same age and sex.

Suggest three possible explanations for this observation.

β carotene may be a protective factor for cancer.

Cancer may give rise to low β carotene levels.

β carotene may need for survival of human

Q2. Table 7 shows results of a cross sectional survey of coronary heart disease (CHD) among male farm workers aged 40 -70 years by their occupational physical activity.

Table 7. Distribution of coronary heart disease (CHD) among male farm workers aged 40 -70 years by their occupational physical activity

Level of physical activity	No. examined	No. with CHD	Prevalence rate	Age-adjusted prevalence rate
Low	89	14	157.3/10^3	126/ 1000
High	90	3	33.3/10^3	36/1000
Total	179	17	9.5	87/ 1000

2.1 Complete column 4 of the table.

2.2 Describe and comment on the findings. workers with low physical activity has a higher tendency to acquire CHD

2.3 Suggest why age-adjusted prevalence rates have been calculated.

To standardize the results

Q. 3 Given below are the findings of a cross sectional survey carried out among a sample of 600 flat dwellers within the age group 35-45 years.

Level of physical	Diabetes mellitus*		Total	
	Absent	Present		
Low	305	20	325	$61.5/10^3$
Medium	171	19	190	$100/10^3$
High	65	20	85	$235.2/10^3$
Total	541	59	600	

* Based on defined criteria

3.1 Describe the findings.

People having high physical activity has higher tendency to get DM.

3.2 What conclusions could you draw?

DM is higher in hyperactive people.

People with DM engage in physical activities than others.

3.3 What additional information would you like to have?

Activity level before diagnosis of Diabetes mellitus

Stress

Q. 3 Given below are the findings of a cross sectional survey carried out among a sample of 600 flat dwellers within the age group 35-45 years.

Level of physical	Diabetes mellitus*		Total	
	Absent	Present		
Low	305	20	325	$61.5/10^3$
Medium	171	19	190	$100/10^3$
High	65	20	85	$235.2/10^3$
Total	541	59	600	

* Based on defined criteria

3.1 Describe the findings.

People having high physical activity has higher tendency to get DM.

3.2 What conclusions could you draw?

DM is higher in hyperactive people.
People with DM engage in physical activities than others.

3.3 What additional information would you like to have?

Activity level before diagnosis of Diabetes mellitus
Stress

Q. 3 Given below are the findings of a cross sectional survey carried out among a sample of 600 flat dwellers within the age group 35-45 years.

Level of physical	Diabetes mellitus*		Total	
	Absent	Present		
Low	305	20	325	$61.5/10^3$
Medium	171	19	190	$100/10^3$
High	65	20	85	$235.2/10^3$
Total	541	59	600	

* Based on defined criteria

3.1 Describe the findings.

People having high physical activity has higher tendency to get DM.

3.2 What conclusions could you draw?

DM is higher in hyperactive people.

1

People with DM engage in physical activities than others.

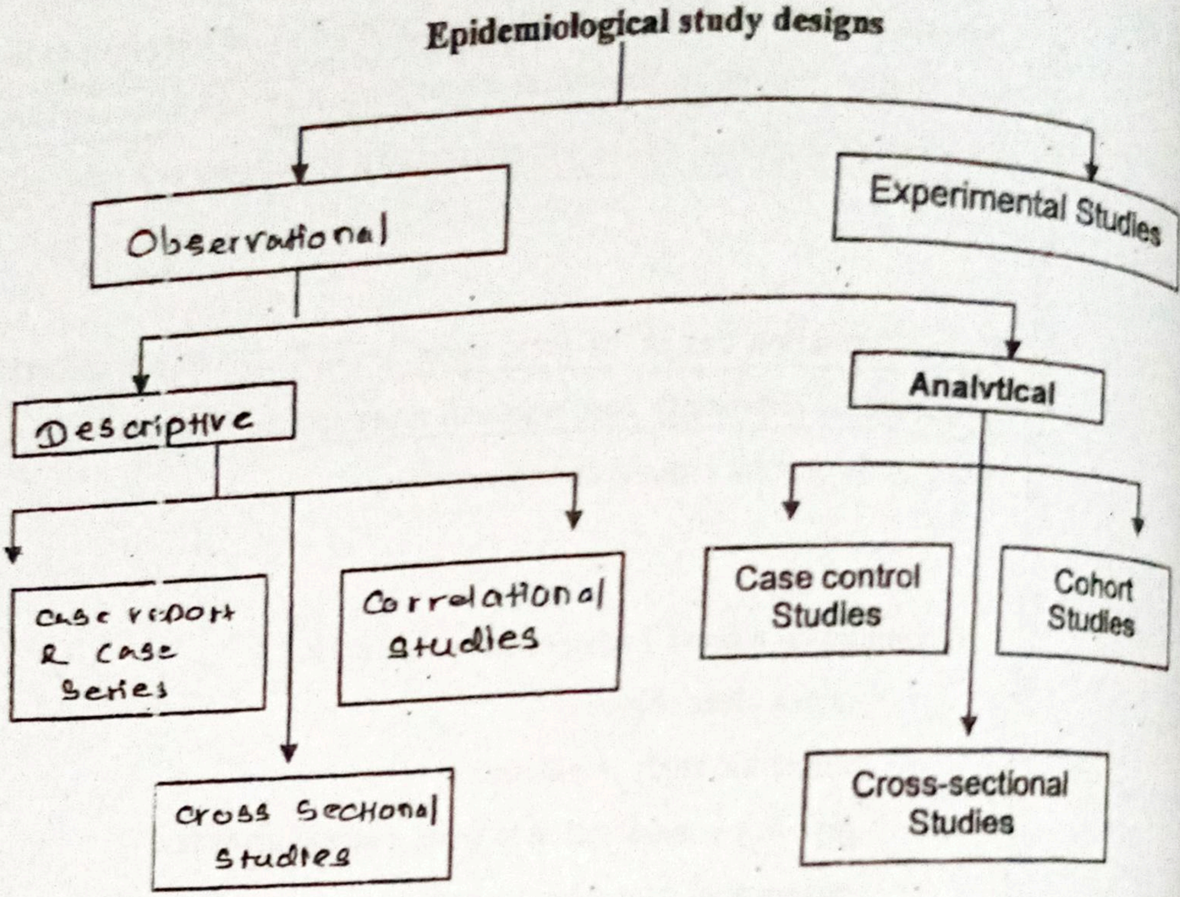
3.3 What additional information would you like to have?

Activity level before diagnosis of Diabetes melitus

Stress

Summary

i. What are the types of study designs you have learnt up to now?



ii. How do they differ from each other?

Observational Studies.	Analytical Studies.
* No hypothesis at the beginning.	* There is a hypothesis at
* Formulation of hypothesis with results.	* Design the the study to hypothesis
* Focus of interest is population or reprensative sample	* Focus of interest is the individual
* No comparison grouns	* comparison groups there.

4. Introduction to analytical studies

Up to now, we have discussed the following epidemiological study designs under observational studies:

4.1.1 Descriptive studies

a. Case reports and case series

b. Correlational studies

c. Cross sectional studies

As you have learnt, descriptive studies are very useful to describe patterns of disease occurrence with respect to person, place and time and thereby formulate hypothesis about a disease and its risk factors.

What is a hypothesis?

It is a statement of belief or intention, which one expects to prove or disprove. This statement relates to certain factor/s which cause/s or relate/s to the occurrence of a disease.

☞ Refer to Exercise 1 in Chapter 3 (pages 26-28). In this cross-sectional survey that has been carried out to describe the prevalence and characteristics of hepatitis B infection, what hypotheses are you able to formulate?

Hint: refer to your answer given in 1.3 in this Exercise

Males are more susceptible for acquiring hep.B infection.

O.A age group is more "

unmarried people are " " " " " "

☞ In the same manner, you can also formulate hypotheses for Q2 and Q3 in Exercise 4 in the same chapter (pages 37-38).

Hint: refer to your answers given in 2.2 and 3.1 in this Exercise

Workers with low physical activity are more prone to get CHD

DM has high prevalence among hyperactive ~~pers~~ persons

Let us now take a look at the conclusions that we drew from these cross-sectional surveys.

For example, in Q3 in Exercise 4, conclusions were:

- High physical activity levels seem to be associated with the prevalence of diabetes among flat dwellers

  Or

- A diagnosis of diabetes may have influenced the flat dwellers to change their behaviour from being less physically active to being more physically active

As you can see, we cannot definitely come to a conclusion that low physical activity is associated with diabetes. This is a weakness in all cross-sectional study designs as they measure both exposure or risk factor and the disease status **at the same time** in each individual. This makes it difficult to determine whether the exposure came before the disease or after the disease.

It is not possible to establish causal or temporal relationships from data collected in a cross sectional time frame.

If we are to establish such relationships about diseases, we need to have a study design with comparison groups either for the exposure or the outcome and work backwards or forwards. This is one of the key features in analytical studies. In the following chapters in this volume, we will study analytical studies - the second type of study designs that come under observational studies.

3.1.2 Analytical studies

In analytical studies, the investigator begins with a hypothesis and designs the study to specifically test that hypothesis. This is different to a descriptive study, which does not begin with a hypothesis but its results used for formulating a hypothesis. The term 'analytical' implies that the study is designed to establish the 'cause' of a disease by looking for **associations** between disease occurrence and its exposure to a risk factor.

The basic approach in analytical studies is to test a specific hypothesis that was formulated based on the findings of a cross-sectional study

Another feature in an analytical study is that subject of interest is the individual in the population. This is different to descriptive studies where the population or a representative sample of that population is the focus of interest. However, as in descriptive studies, the inferences are made to the population from which the individuals are drawn from, and not only to the individuals who are actually studied.

☞ **It is important to note that although analytical studies help in testing a hypothesis about associations between disease occurrence and its exposure to risk factors, this alone will in no way imply 'causality' of a disease.** Such a judgement of causality can only be made by taking into account all evidence that is available according to a set of criteria. These criteria that assist in judging causality of a disease include: strength of the association, biological credibility of the hypothesis, consistency of the findings, temporal sequence and the presence of a dose- response relationship.

Analytical studies have the following features:

- They are used to test hypotheses
- There is always a comparison group
- The basis for identifying the two groups is:
  - presence/ absence of exposure
  - presence/ absence of disease

MCQ

There are three types of analytical studies. They are:

a. Cohort Study

b. Case-control Study

c. Cross sectional design in analytical study

5. Cohort studies

5.1 Types of cohort studies

A major type of analytical study designs is the cohort study. This design is also called **follow up or incidence studies.**

'Cohort' was the Roman term for a group of soldiers that marched together. In clinical research, cohort is a group of people grouped together with a common cause or a group of persons with a common characteristic or experience. For example, a group of persons born during the same year makes a birth cohort.

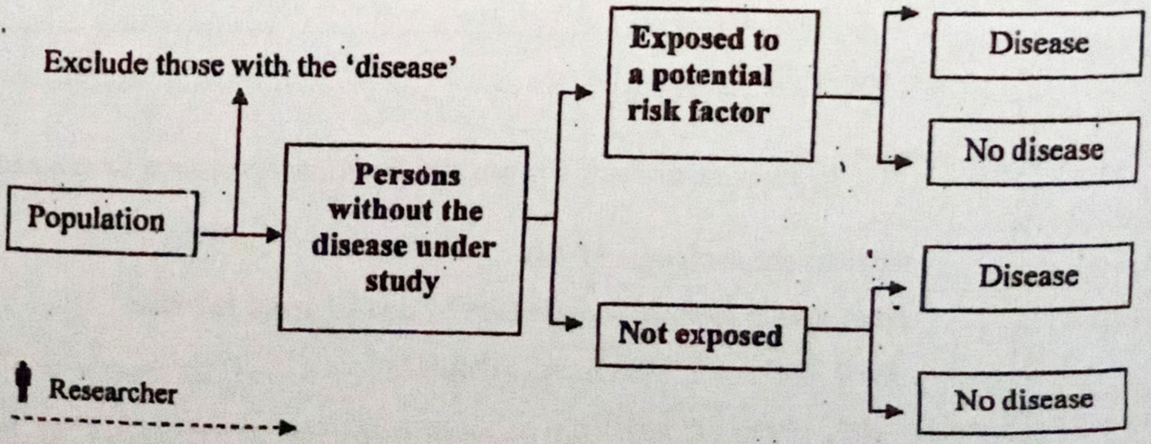
A cohort is a group of persons who share a common characteristic or an experience

Let us now consider a group of people who are free of a particular 'disease' under study. This group can be further divided into two groups based on their exposure to a 'potential cause' of this disease i.e.

- Group with the exposure to a potential cause
- Group without the exposure to a potential cause

In epidemiology, we refer to this particular 'disease' as the **outcome** and the 'potential cause' as the **risk factor** or **exposure.** Both exposed and non-exposed groups are then followed up to see whether individuals in each group would develop the outcome. This sequence of events makes up a 'cohort study' and is illustrated in Figure 1.

Figure 1. Sequence of events in a cohort study design



In a cohort study, the direction of inquiry is from exposure to disease.

	Direction of inquiry	
Exposure	$\longrightarrow$	**Disease**

Since cohort studies are interested in exposures *preceding* the disease outcomes, these studies can establish the time sequence of events and thereby strengthen the inferences made about exposures as 'potential causes' of disease outcomes.

There are two types of cohort studies. Of the two, one is called **prospective cohort studies**, in which the investigator is looking **forwards** from potential cause (exposure) to possible disease (outcome).

Given below is a classic example of a prospective cohort study.

Prospective Cohort Study

The Nurses' Health Study examined the incidence and risk factors for common diseases in women. The basic steps in performing the study were to

1. *Assemble the cohort.* In 1976 the investigators obtained lists of registered nurses aged 25 to 42 in the 11 most populous states and mailed them an invitation to participate in the study.
2. *Measure exposures (potential risk factors).* They mailed a questionnaire to obtain information about their diet and received completed questionnaires from 121,700 nurses. They sent questionnaires every 2 years for the next two decades and updated the status of their diet measured at baseline.
3. *Follow up the cohort and measure outcomes.* The periodic questionnaires also included questions about the occurrence of a variety of disease outcomes, which were then confirmed by review of medical records.

The prospective approach allowed investigators to measure exposures on diet completely and accurately at baseline. The cohort design allowed them to collect data on subsequent outcomes. The large size of the cohort and extended period of follow up have provided an unparalleled opportunity to study risk factors for various forms of heart disease, cancer, and other common diseases. For example, the investigators examined the hypothesis that high intake of dietary fibre is associated with a decreased risk of colorectal cancer. Fibre intake was assessed in 1980, and 787 cases of colon cancer were confirmed between 1980 and 1994. The rate of colon cancer among women in the lowest decile of dietary fibre intake was similar to the rate in women in the highest decile of fibre intake (relative risk = 1.0; 95% confidence interval, 0.7 to 1.4). The investigators also adjusted the analysis for potential confounding factors, and this did not change the result. The large number of cases of colon cancer and the quality of the methods support the conclusion that high intake of dietary fibre does not prevent colon cancer.

Ref: Fucsh CS, Giovannucci EL, Colditz GA, *et al.* Dietary fibre and the risk of colo-rectal cancer and adenoma in women. *N Engl J Med* 1999;340:169-76.

1. What were the exposed and non-exposed groups and the disease outcome/s that were studied by the researchers in this study?

Exposed group - ~~th~~ Nurses having high fibre intake.

Non-exposed groups - Nurses with low fibre dietary intake

outcome - colon cancer

2. What was the assumption that the researchers made about the cohort of nurses before they were enrolled into the study?

Researchers ~~Hiah~~ Considered the nurses as normal women. no outcome which is being studying

3. Can you think of any difficulties the researchers faced in carrying out this study? - There is year ... up is longer ...

Some people die, some may change residence, missing data, completely & accurately. spending

Several questionnaires to be made - cost.

Reliability of information. with long term follow up people can be ~~missed~~ lost

There is another type of cohort studies, in which the investigator is looking **backwards** from potential cause (exposure) to possible disease (outcome). These are called **retrospective cohort studies.**

Given below is a classic example of a retrospective cohort study.

Retrospective Cohort Study

To describe the effect of asbestos dust exposure on the mortality of cancers, Enterline analyzed data in a retrospective cohort study among asbestos workers. The basic steps in performing the study were to

1. *Identify a suitable cohort.* The investigators used the social security tax returns filed with the United States Bureau of Internal Revenue during 1948 to 1951 to identify asbestos workers who retired normally at age 65; those who retired before age 65 for personal reasons but lived to age 65; and men who retired before age 65 because of a disability but who lived to be 65.

2. *Collect data about exposures.* There was a total of 1 376 men in this study who reached age 65. Of them, complete exposure and job histories were available for 1 348. Dust levels at each job site and time period of exposure were expressed as million particles per cubic foot of air (mppcf).

3. *Collect data about subsequent outcomes that occurred at a later time.* They collected data on 58 cancer deaths occurring in this group between 1948 and 1963 from claims filed with the Social Security administration and the corresponding death certificates obtained from state health departments.

The observed mortality among these men was compared with an expected mortality of the entire US white male population living in the time-age intervals that characterized the retired population. The investigators found that men who retired from an industry with asbestos dust exposure had an overall mortality rate 14.7% higher than all US males. This excess was due almost entirely to cancer and respiratory diseases. For cancer, the greatest excess was in respiratory system cancers. For respiratory diseases, the excess was entirely due to pneumoconiosis and pulmonary fibrosis.

Ref: Enterline PE. Mortality among asbestos product workers in the United States. ***Ann N Y Acad Sci*** **1965;132:156**

1. What were the exposed and non-exposed groups and the disease outcome/s that were studied by the researchers in this study?

Ar Retired asbestos workers ~~retired~~ at age 65.
Exposed group - People who ~~exposed~~ to

Non exposed group - White male population for in same time - age interval.

Outcome - respiratory system cancer & diseases.

2. How did the researchers ensure that the exposures preceded the outcomes?

Studying that asbestos dust exposure has overall mortality rate 19.7?
of all US males.

3. If you re-design the above study into a prospective cohort study, can you think of any advantages and disadvantages that you may have?

Advantages
Data are more reliable
More access to information

Disadvantages.
More time consuming

It should be clear to you that in both types of cohort studies, the study begins with exposed and non-exposed groups. However, in retrospective cohort studies, all relevant events (both exposure to potential risk factor and disease outcome) *have already occurred* when the study is initiated. In contrast, in prospective studies, the exposures have occurred at the time the study is begun but the disease outcomes have *certainly not yet occurred* and therefore, the participants need to be followed up into the future to assess the incidence rates of the disease outcome.

Terms 'prospective' and 'retrospective'
are used only in relation to the timing of data collection and
not to refer to a particular study design type.

1. Indentify a suitable
2. Select an appropriate group.
3. Collect data on exposure nt outcome
4. approaches to follow
5. Analysis & interpretation

5.2 Essential steps in carrying out a cohort study

There are five essential steps in carrying out a cohort study.

a. Identify a suitable cohort

For a common exposure such as cigarette smoking and betel chewing, a large number of exposed persons could be identified from the general population (community cohort). For rare exposures such as certain occupations (e.g. asbestos, cinnamon industry), medical therapy or procedures (e.g. chemotherapy, x-ray treatment), environmental hazards (e.g. high tension wires, Atomic bomb in Hiroshima), it is necessary to identify a group who have undergone that specific exposure or experience. Advantages of such cohorts include:

- It allows collection of a sufficient number of exposed persons within a short time
- It leads to identification of aetiological agents in special circumstances
- It allows more complete and accurate information on exposures and good compliance at follow up. For example, Doll and Hill (1950) utilised British Doctors as the cohort for studying smoking and lung cancer because of their ability in providing accurate information about their smoking habits.

b. Select an appropriate comparison group

The choice of a non-exposed group is crucial and difficult. Ideally, the non-exposed group needs to be as similar as possible to the exposed group with respect to all other factors that are related to the disease except the exposure under study.

In a single general cohort, an **internal** comparison group can be utilised. For example, experience of the cohort members classified as having an exposure is compared with that of members of the same cohort who are either non-exposed or exposed to different degrees. In contrast, in some occupational settings, there might not be a comparison group that could be definitely identified as non-exposed. In this instance, an **external** comparison group could be utilised. For example, in a study assessing the effect of quarry dust on respiratory diseases, an appropriate comparison group for a cohort of quarry workers was selected from hospital labourers. This study was further strengthened by including **multiple** comparison groups from different occupational settings (e.g. cotton textile workers, rubber tappers). This is very useful especially when no single group is comparable with the exposed group.

c. Collect data on exposure and subsequent outcomes

Data on basic characteristics of the cohort, exposure and subsequent disease outcomes need to be collected. Information on exposure status could be obtained from a number of sources. Some of these include: surveys with follow up procedures/ interviews, medical and employment records monitored over time and periodic medical examinations and interviews.

The use of pre-existing records offers a number of advantages as well as disadvantages.

State one advantage and disadvantage in using pre-existing records.

DisAdvantage - Less time consuming. Easy to access

Disadvantage - Lack of reliability

For a cohort study with fatal endpoints, outcome data can be obtained from death certificates. It is the most reliable method for assessing all-cause mortality as the outcome in a study but not so much for mortality from a specific disease. For non-fatal end points, outcome data can be obtained from physician's records, BHT, hospital registers, etc. Additional information such as pathology reports, hospital records can be used to confirm the diagnosis.

Whatever method is used for identifying disease outcomes, they should be equally applied to both exposed and non-exposed groups.

d. Approaches to follow up

Whether prospective or retrospective, in any cohort study, the collection of outcome data involves tracing or following up the cohort members from the point of exposure to the point of disease outcome. This poses the biggest challenge in conducting a cohort study. It is also the main reason for its high cost in terms of money and time. In general, the longer the duration of follow up, the more difficult it will be to achieve complete follow up because the cohort members are more likely to move, to change jobs, to change names, etc.

In a cohort study, disease outcome data is collected from both exposed and non-exposed groups via interviews, questionnaires and examination of records. Since this information is very crucial in drawing conclusions about associations, we need to ensure that we do not introduce any bias during data collection.

What is bias?

It is an error made in epidemiological studies due to known sources of variation resulting in an incorrect estimation of the association between an exposure and the risk of disease. Usually, this error will distort study findings in one direction.

They are mainly of three types due to differences in the way:

- study subjects are selected into a study (**selection bias**)
- information is reported by the study subjects (**recall bias**) or obtained or interpreted by the researchers (**interviewer bias**) or loss of participants to follow up (**loss to follow up**)

In a cohort study, bias can be introduced mainly due to loss-to-follow up. Since both exposure and outcome have occurred at the beginning of a study, recall bias could influence the classification of exposures in retrospective cohort studies. Since exposure is measured prior to the disease outcome, it is unlikely that such bias could influence the classification of exposures in prospective cohort studies. However, interviewer bias could be introduced while measuring disease outcomes in both types of cohort studies.

c. Analysis and interpretation

Once the data is collected, the relationship between the exposure variable and outcome can be presented in a two by two table, as shown below.

Table 1. Relationship between the exposure and outcome in a cohort study

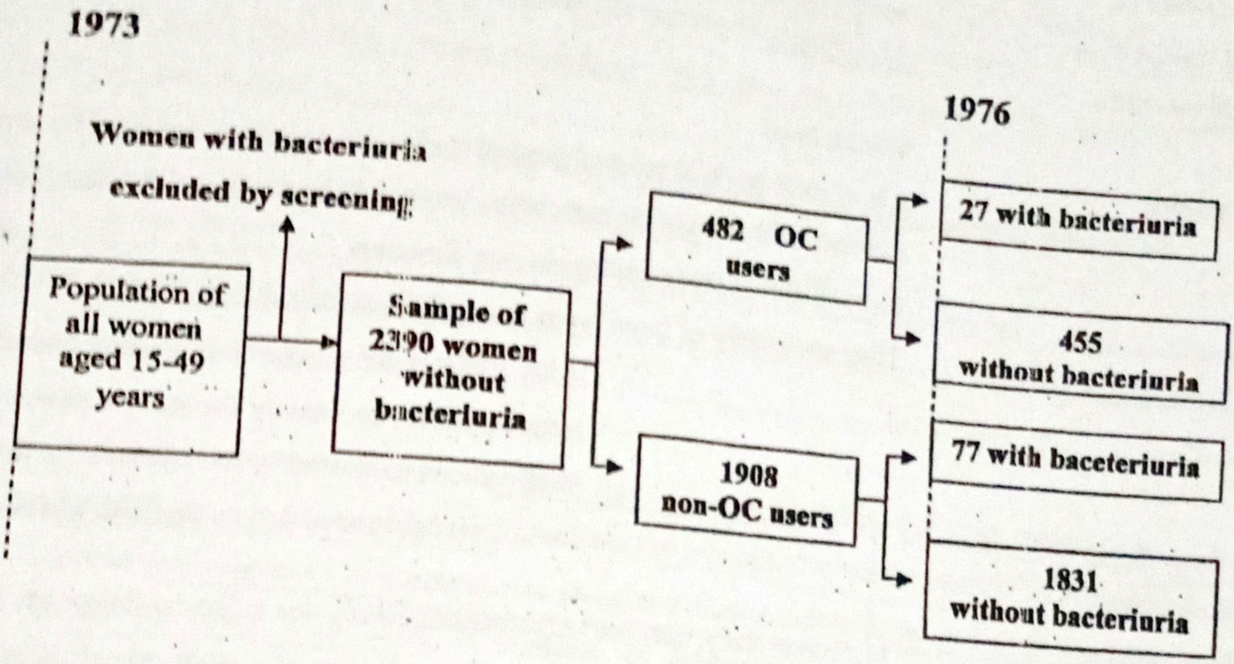
Exposure status	**Outcome status**		**Total**
	Present	Absent	
Yes	a	b	a + b
No	c	d	c + d
Total	a + c	b + d	N

N = a+b+c+d

Exercise 1

Q1.

Figure 2. A cohort study to investigate the relationship between the use of oral contraceptives (OC) and bacteriuria among women aged 15-49 years



Using information in Figure 2,

1.1 Write the 2x2 table to demonstrate the relationship between OC use and bacteriuria.

Exposure State OCP	Outcome Status Present	absent	Total
Yes	27	455	482
No	77	1831	1908
Total	104	2286	2390

1.2 Write the hypothesis for the association that needs to be tested in this cohort study.

~~"using of OCP" causes bacteriuria.~~

There is no association between ~~using of~~ OCP use & bacteriuria

1.3 The above hypothesis reworded to read as a null hypothesis, will say that there is no association b/w use of OCs & bacteriuria. To test whether ē above null hypothesis is valid a stastical test has to be performed. This test will test for the difference of the following proportions of OC users &

$P_1 =$ 27/48... ... = 0.04

The test to be performed is called ē SND Community Stream, FM, Colombo
test or z test & will be performed at a 5% level

Expression is as follows $SND = \frac{P_1 - P_2}{SE_{P_1 - P_2}}$

1.3 Test the given hypothesis.

~~Incidence~~ Percentage of disease of exposed people $= \frac{27}{482} \times 100\% = 5.6\%$

~~Incidence~~ percentage of disease of non exposed people $= \frac{77}{1908} \times 100\% = 4$

SND = Z

$$Z = \frac{P_1 - P_2 - 0}{SE_{P_1 - P_2}}$$

$$SE_{P_1 - P_2} = \sqrt{\frac{P_1 q_1}{n_1} + \frac{P_2 q_2}{n_2}} = \sqrt{\frac{5.6 \times 94.4}{482} + \frac{4 \times 96}{1908}}$$

$$= \sqrt{1.09 + 0.2} = \sqrt{1.29}$$

$q_1 = (100 - P_1)$

$$= 1.13$$

$$Z = \frac{5.6 - 4}{1.13} = \frac{1.6}{1.13} = 1.41$$

Null hypothesis - There is no association between OCP use & bacteriuria.

5% confidence interval - 1.96

As SDN value < 1.96

∴ Null hypothesis can not be rejected

i.e There is no association between OCP & bacteriuria.

What you have done in 1.3 is testing a hypothesis for a cohort study using a statistical approach. Let us now consider testing the same hypothesis using an epidemiological approach.

By substituting to the above expression SND = 0.016 / 0.012 = 1.33

Since 1.33 is < 1.96, the null hypothesis is accepted &

can not be rejected. That means there is no association b/w the two proportions P_1 & P_2 or OC use & development of bacteriuria.

In cross sectional studies... measure of disease frequency. Prevalence.

5.3 Relative Risk

Name the measures of disease frequency that may be derived from a cohort study.

Incidence

We make use of this measure of disease frequency to assess the effect of an exposure on a disease. It is by calculating a ratio of the incidence rates of a disease in the exposed and the non-exposed groups. This ratio is called the **relative risk (RR) or risk ratio.**

If the incidence of disease in the exposed group is denoted by I_1 and the incidence of disease in the non-exposed group by **Io**, this is how we can write the RR:

$$\text{Relative risk (RR)} = \frac{I_1}{Io}$$

I_1 = Incidence of disease in exposed group

I_0 = Incidence of diseases in non exposed group.

- **Calculation of RR**

i. For a cohort study with **count data,** the relative risk is calculated as the ratio of the **cumulative incidence** of a disease in the exposed and the non-exposed groups.

$$\text{Relative risk (RR)} = \frac{\text{Cumulative incidence rate of disease in the exposed group } (CI_1)}{\text{Cumulative incidence rate of disease in the non-exposed group } (CI_o)}$$

Complete the following equation for relative risk using symbols given in Table 1.

$$RR = \frac{a/(a+b)}{c/(c+d)}$$

1.4 Calculate the RR for Q1 in Exercise 1.

Relative risk = 27/482 / 77/1908 = 0.056 / 0.040 = 1.4

In PG exam

define these words. I_0, I_1

Name the exposure = Risk factor. & outcome

ii. Let us consider a cohort study with person-time follow-up data, as shown in Table 2.

Table 2. Presentation of data from a cohort study with person time data

Exposure status	**Outcome status**		**Person time units**
	Present	Absent	
Yes	a	-	PY_1
No	c	-	PY_0
Total	a + c		$PY_1 + PY_0$

Please refer to volume 1 pages 13-14 to understand person-time data.

For a cohort study with **person-time data**, the relative risk is calculated as the ratio of the **incidence density** of a disease in the exposed group and the non-exposed group.

$$\text{Relative risk (RR)} = \frac{\text{Incidence density of disease in the exposed group } (ID_1)}{\text{Incidence density of disease in the non-exposed group } (ID_0)}$$

Complete the following equation for relative risk using symbols given in Table 2.

$$RR = \frac{a/(\ldots PY_1 \ldots)}{c/(\ldots PY_0 \ldots)}$$

Q2. A cohort study was carried out among postmenopausal female nurses to investigate the relationship between postmenopausal hormone use and coronary heart disease (CHD). After a total of 54,308.7 person-years of follow-up, 30 women who reported use of post menopausal hormones developed CHD. Among the non-hormone users, 60 developed CHD after 51,477.5 person-years of follow-up.

2.1 Write the 2x2 table for the above data.

Exposure status	Outcome status		Person time units
	Present	Absent	
Yes	30		54 308.7
No	60		51 477.5
Total			[illegible]

2.2 Calculate the RR.

Relative risk $= \dfrac{30/54308.7}{60/51477.5} = \dfrac{0.0005}{0.0014} = 0.5$

~~0.47~~.

- **Interpretation of relative risk**

Relative risk estimates the **magnitude or strength of an association** between an exposure and a disease. It indicates the likelihood of developing a disease in a group of people exposed to a potential risk factor relative to the non-exposed group and therefore assessing the likelihood of an association representing a causal relationship.

RR is an indicator of the **strength** of an association between an exposure and a disease.

Let us now see how the numerical value of RR could be interpreted.

- **A relative risk of 1.0** indicates that the incidence rates of disease in the exposed and non-exposed are identical. In other words, the likelihood of developing a disease in a group of people exposed to a risk factor relative to a group not exposed to it is equal.

If the relative risk is 1, there is **no association** between the exposure and the disease outcome.

- **A relative risk greater than 1** indicates that the incidence rate of disease in the exposed is higher than that of the non-exposed group. In other words, the likelihood of developing a disease in a group of people exposed to a risk factor is higher than that in a group not exposed to it.

If the relative risk is > 1, there is a **positive association** or an **increased** risk of disease among those exposed to a risk factor compared to the non-exposed.

Using this knowledge,

1.5 How do you interpret the RR that you obtained for Q1 in Exercise 1?

O C users were 1.4 times more likely to develop bacteriuria compared to non OC ~~ea~~ users

We say that OC users had **1.4 times the risk** of developing bacteriuria compared to non-OC users or OC users were **1.4 times more likely** to develop bacteriuria compared to non-OC users.

How would you interpret a RR less than 1?

There is a decreased risk of disease among those who exposed risk factor compared to the non - exposed

We have obtained a relative risk of 0.5 in the study given in Q2 in Exercise 1. It indicates that the women who used post-menopausal hormones had only half or 0.5 times the risk of developing CHD compared with non- post-menopausal hormone users or women who used post-menopausal hormones were only 0.5 times more likely to develop CHD compared to non-hormone users.

5.4 Attributable Risk

Another way of testing a hypothesis in a cohort study using an epidemiological approach is by comparing the difference in disease occurrence between the exposed and non-exposed groups. This difference is called the **attributable risk (AR)**, **excess risk** or **absolute risk.** It is the difference in the incidence rates of disease between the exposed and non-exposed groups.

This is how we write the AR.

Attributable risk (AR) = Incidence in the exposed (I_I) - Incidence in the non-exposed (I_0)

- **Calculation of AR**

i. In a cohort study with **count data**, the AR is calculated as the difference of the **cumulative incidence rates** between the exposed and the non-exposed groups. Complete the following equation for attributable risk using symbols given in Table 1.

$$AR = (a / \text{....}a+b\text{....}) - (c / \text{....}c+d\text{....})$$

ii. In a cohort study with **person-time data**, AR is calculated as the difference of the **incident densities** between the exposed and the non-exposed groups.

$$AR = (a / \text{....}PY_1\text{....}) - (c / \text{....}PY_0\text{....})$$

1.6 Calculate the AR for Q1 in Exercise 1.

AR for Q1 = $\frac{27}{482} - \frac{77}{1908}$ = 5.6 − 4

= 1.6 per 100 people

incidence

Risk of getting bacteriuria

- **Interpretation of attributable risk**

Attributable risk estimates the **absolute effect** of an exposure or the **excess risk** of a disease in the exposed group compared to the non-exposed groups. The AR is therefore used to quantify the risk of disease in the exposed group that is **attributable** to the exposure itself after excluding the risks due to all other potential causes other than the exposure under study that could have led to the disease outcome.

AR is an indicator of the risk of a particular disease **attributable** to the exposure itself.

It is noteworthy that the interpretation of AR is valid only if a cause-effect relationship exists between the exposure and disease under study.

Let us now see how the numerical value of AR could be interpreted.

- An **attributable risk of 0** indicates that there is no difference in the incidence rates of disease between the exposed and the non-exposed groups.

If the attributable risk is 0, there is no **association** between the exposure and the disease outcome.

- If there is a causal association between the exposure and the disease, an **attributable risk greater than 0** indicates the incidence rate of a disease in the exposed that can be attributed to the exposure itself. Alternatively, it indicates the incidence rate of a disease in the exposed that could be eliminated if the exposure was removed.

If the attributable risk is > 0, there is an **excess risk** of disease among the exposed that can be attributable to that exposure.

Using this knowledge,

1.7 How do you interpret the AR that you obtained for Q1 in Exercise 1?

There is an ~~increased~~ excess risk of disease among the exposed that can be attributable to that exposure. The excess incidence of bacteriuria among OC users that can be attributed to their OC use is 1.6 per 100 women.

We say that, if there is a causal association between OC use and bacteriruia, the **excess incidence** of bacteriruria among the OC users that can be attributed to their OC use is 1.6 per 100 women.

- **What is the importance of RR and AR for a medical person?**

- RR and AR are useful in measuring the associations between exposures and disease outcomes and thereby assessing the risk of developing a disease in a population.

RR and AR are called 'measures of associations'.

It is important to note that although both RR and AR are measures of associations, these provide very different information about the risk factors of diseases:

- RR provides information about the strength of an association between an exposure and a disease outcome
- Assuming that there is a causal relationship between an exposure (risk factor) and the outcome (disease), AR provides information about the extent of the public health impact, if the exposure is removed from the population-at risk

Example

In the classic study that was carried out among British male physicians to assess the relationship of cigarette smoking (exposure) with lung cancer and CHD deaths (outcomes), the following results were obtained:

For lung cancer: RR = 14; AR = $130/10^5$/year

For CHD: RR = 1.6; AR = $256/10^5$/year

Ref: *Doll R and Peto AB. A study on the aetiology of carcinoma of the lung, 1952.Br Med J. 2:127.*

Risk of developing lung cancer is among smokers is 14 times higher than ~~in~~ among non smokers

Accordingly, we can conclude that cigarette smoking is a much stronger risk factor for dying of lung cancer compared to CHD. However, if smoking is causally related to both diseases, prevention of cigarette smoking in the population would prevent far more deaths from CHD than from lung cancer among the smokers in the population. Therefore, the public health impact of preventing smoking in the community will be far greater for CHD than for lung cancer.

1. What measure of association is useful for an epidemiologist? .... Relative risk
2. What measure of association is useful for a public health professional? .... attributable
want to prevent death & having disease.

5.5 Advantages and disadvantages of cohort studies

Advantages:

- Provides a complete description of experiences *subsequent* to an exposure such as incidence rates and natural history of diseases and their staging.
- Provides information on more than one disease related to the same risk factor.
- Helpful for determining whether there is a cause - effect relationship since the risk factor precedes in time the occurrence of disease.
- Both relative and attributable risks can be calculated.

Disadvantages:

- Cohort studies are long term and therefore are very expensive, time consuming and not always feasible.
- Large numbers need to be followed up and are unsuitable for studying rare diseases. It is usually difficult to find and manage samples of this size.
- The most serious problem encountered is loss to follow-up of participants or interviewers. This can affect the validity of the conclusions and may make the samples less representative of their source population.
- Over the period of observations, there may be changes in the exposure status of the subjects. Many changes may occur, which may influence the relationship between exposure and disease and may confuse the issue of association.
- The study itself may influence the behaviour of persons under investigation in such a way that it may influence the development of the disease.
- Serious ethical issues may arise with apparent disease excess before data completion.

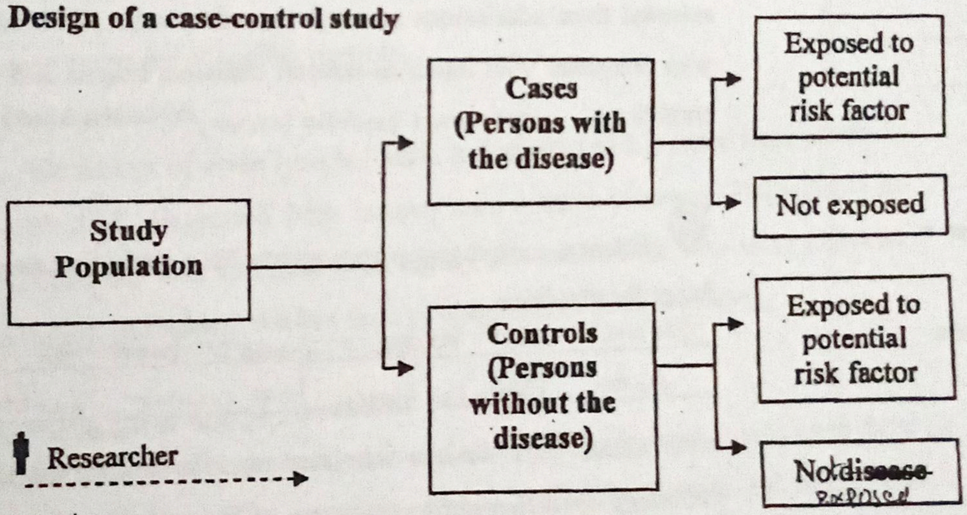
6. Case-control studies

A case control study is the second type of analytical studies under observational studies. In this study design, we divide the subjects into two groups based on the presence or absence of a disease following an exposure to a potential risk factor i.e.

- Group with the disease outcome
- Group without the disease outcome

Let us now see how a case-control study is designed. As illustrated in Figure 1, subjects are selected to the study on the basis of whether they do (cases) or do not (controls) have the disease of interest. History of exposure is then obtained from both cases and controls. Finally, the association between the exposure and the outcome is studied by comparing the exposure pattern of the cases with the exposure pattern seen in the controls.

Figure 1: Design of a case-control study



As you can see, unlike a cohort study that begins with exposed and non-exposed groups and proceeds to assess their outcomes, a case-control study begins with two groups of cases and controls and proceeds to assess their exposures. Therefore, essentially in case-control studies, all relevant events (both exposure to potential risk factor and disease outcome) *have already occurred* when the study is initiated.

This study design is also called 'retrospective study' since the investigator is looking **backward** from disease outcomes to possible exposures. But this term should not be used to refer to a case-control study. As we learnt in chapter 2, prospective and retrospective terms should be reserved only to refer to the timing of data collection of any study.

Given below is a classic example of a case-control study.

Case-Control Study

Since intramuscular (IM) vitamin K is given routinely to newborns in the United States, a pair of studies reporting a doubling in the risk of childhood cancer among those who had received IM vitamin K caused quite a stir. To investigate this association further, German investigators carried out a case-control study. The basic steps in performing this study were to:

1. **Select the sample of cases**—107 children with leukaemia from the German Childhood Cancer Registry were selected.

2. **Select the sample of controls**—107 children matched by sex and date of birth and randomly selected from children living in the same town as the case at the time of diagnosis (from local government residential registration records) were selected.

3. **Measure the predictor variable**—reviewed medical records to determine which cases and controls had received intramuscular vitamin K in the newborn period.

The authors found that 69 of 107 cases (64%) and 63 of 107 controls (59%) had been exposed to IM vitamin K, for an odds ratio of 1.2 (95% confidence interval [CI], 0.7 to 2.3). (See Appendix 8.A for the calculation.) Thus this study did not confirm the existence of an association between the receipt of IM vitamin K as a newborn and subsequent childhood leukemia, although the point estimate and upper limit of the 95% CI leave open the possibility of a clinically important increase in leukemia.*

Ref: Von Kries R, Gobel U, Hachmeister A, Kaletsch U, Michaelis J. Vitamin K and childhood cancer: a population based case conrol study in Lower Saxony, Germany.

6.1 Essential steps in carrying out a case-control study

There are 4 essential steps in carrying out a case-control study. They are,

a. Select cases and controls
b. Match cases and controls
c. Measure the exposure status
d. Analysis and interpretation

a. Select cases and controls

Comparability of cases and controls is essential in a case-control study, which would enable the researcher to conclude that the disease among the cases was most likely due to the exposure under study. Cases and controls should be comparable by their baseline risk of developing the disease other than from the exposure under study and also by the accuracy and completeness of exposure data.

- **Cases**

Cases should represent a disease entity as homogenous as possible. For example, hepatitis A and B have very different aetiologies and it would be wrong to consider all 'hepatitis infections' as cases in a study. Therefore, it is important to identify cases using a case definition.

Case definition

Often in practice, individuals who fit into the definition of a case, in a particular setting, identified within a specific period of time are included as cases. Relevant information about the diagnosis can be collected using questionnaires, direct questioning and existing records such as bed head tickets, diagnosis cards and investigation reports.

A case definition should be clear and unambiguous with criteria for inclusion and exclusion from the study

Develop a case definition for obesity.

more than 30kg/m²

Individuals who are having BMI value ~~which~~ belongs to ~~overweigh'~~ obesity category

Sources of cases

Cases can be selected from a number of sources. Such sources may include:

- cases attending a clinic or admitted to/discharged from a hospital within a specified period of time
- cases identified in a community survey or surveillance program at a specified time

Whatever the source that you may use, 'newly diagnosed' cases are preferred to 'prevalent' cases for some of the reasons given below:

- Prevalent cases can have factors that would have enabled them to survive, and not necessarily the risk factors. Incident cases have risk factors at the time of diagnosis.
- Prevalent cases will give false information about behaviours related to the advice given, and not really the type of behaviour they have had at the time of diagnosis.

Can you think of a disease for which obtaining prevalent cases is unavoidable?

Congenital malformations. rare diseases

- **Controls**

The crucial and most difficult task in the design of a case-control study is selection of an appropriate control group.

Control definition

Controls should be persons who would have been identified as 'cases' had they not developed the disease. Therefore, any exclusions made in the identification of cases should also be applied equally to the controls.

Sources of controls

There is no control group that is optimal for all situations and may differ according to the scenario. Given below are some examples for different sources of controls.

- **Hospital controls**

In evaluating the association of cigarette smoking with myocardial infarction, cases were identified from admissions to coronary care units of selected hospitals. Controls were selected from admissions to surgical, orthopaedic and medical wards of the same hospital who presented with musculo-skeletal diseases, trauma and a variety of other non-coronary conditions. community population, friends, neighbourhood, relatives.

Select a control group appropriate for a study that assesses physical exertion as a risk factor for abortions. -

Women who ~~have given birth to~~ delivered live babies, during same PUO, same last regular menstrual periods (more or less similar) (2 weeks difference allowed)

Hint: Cases were women who had an abortion by 20 weeks of gestation and with a pathology specimen analyzed by one of the pathology departments in a health area over a given period of time.

Advantages of having hospital controls include:

- They are easily identified; readily available; and more willing than those at home.
- They are more likely than healthy controls to recall exposure events in the past because they are hospitalized and ill and therefore comparable to cases.
- When controls are identified from the same hospital as cases, they have had the same selection factors that influenced the cases to come to that particular hospital.

One disadvantage of having hospital controls is that the disease for which the controls are hospitalized may be associated with the risk factor under study. For example, including patients with bronchitis and pneumonia as controls in a study assessing the relationship between cigarette smoking and lung cancer would underestimate this relationship. It is because smoking is a risk factor for all three diseases and therefore different from the smoking habit of healthy population.

Healthy controls are selected when it is not desirable or feasible to select controls from hospitalized populations. However, they may pose difficulties such as:

- It is often expensive and time consuming.
- They are usually busy; are difficult to meet; and are less motivated to participate.
- Those who volunteer to participate may be very different to the general population.
- They may not recall exposure with the same level of accuracy as hospital controls.

Given below are some examples of healthy controls.

- **General population controls**

In evaluating the risk factors of acute lymphoblastic leukaemia (ALL), cases were children aged 0-9 years with ALL diagnosed from tertiary care centres between (1980-1993) in Québec, Canada. Controls were children from family allowance files who were matched for age, sex, and region of residence at the time of diagnosis of matched cases.

- **Controls from the same source population of cases**

In evaluating the association of electro-magnetic radiation with leukaemia, cases were those who worked in an electric utility company whose underlying cause of death was leukaemia. Controls were selected from workers of the same company who were alive on the date of death of the index case. History of exposure to electro-magnetic radiation was determined for both groups using company job history information.

- **Family, friends or neighbourhood controls**

In evaluating the association of menstrual cycle pattern with endometriosis, cases were women aged 15-49 years who were newly diagnosed patients of endometriosis confirmed by laparoscopy and attending a specialist clinic during a specified period. Each woman was asked to provide names of four friends who were not biologically related; were not patients of the particular specialist clinic; were not known to be having endometriosis; and were within two years of their own age. From this pool of friends, controls were randomly selected.

Family, friends or neighbourhood controls may be more co-operative than other controls because of their interest in the cases and may also offer a degree of similarity in relation to their lifestyles, ethnicity, socio economic status and environment.

Think of an exposure for which family members or friends are also likely to be exposed as the cases, so that it leads to an under-estimation of the true effect.

Food habits, socio economic background, smoking.

b. Match cases and controls

Many risk factors and diseases are related to age, sex, etc. Study results of a case-control study may not be meaningful if the two groups are selected differently in these variables. For example, in a study of exercise and risk of myocardial infarction, factors such as age, sex and smoking are also associated with MI. If we include more smokers as cases in this study, it could lead to an over-estimation of this association. A simple method that eliminates this problem is matching cases and controls.

1. In a case-control study that evaluates the association of exercise with myocardial infarction, what other factors are more likely to distort the findings of this study?

Genetic predisposition, lifestyle, dietary habits.

2. Describe how you would overcome this by carrying out matching in this situation.

Finding cases & controls similar to above conditions.

c. Measure the exposure status

In a case-control study, exposure data is collected from both cases and controls via interviews, questionnaires and examination of records. Since this information is very crucial in drawing conclusions about associations, we need to ensure that we do not introduce any **bias** during data collection (Refer bias in page 10).

Bias can be introduced into a case-control study either by the participants or by the researcher due to differences in the way:

- cases and controls are selected from different settings (**Selection bias**)
- cases and controls recall exposure information (**recall bias**) or

interviewers report or interpret information (**interviewer bias**)

Given below are some precautions that must be taken to minimize such bias:

- Exposure should be measured using an objective method such as well-standardised survey methods, questionnaires and procedures
- Questionnaires and procedures that are used to gather exposure information should be uniformly applied to both cases and controls.
- Neither case nor control should be aware of the hypothesis under study as it might bias the answers given by the participants.

d. Analysis and interpretation

Once the data is collected, like in cohort studies, the relationship between the exposure variable and outcome can be presented in a two by two table, as shown below.

Table 1. Relationship between the exposure and outcome in a cohort study

Exposure status	Outcome status		Total
	Present	Absent	
Yes	a	b	a + b
No	c	d	c + d
Total	a + c	b + d	N*

*N = a+b+c+d

Exercise 1

In a case-control study investigating the association between cigarette smoking and lung cancer, it was found that of the 518 cases of lung cancer 499 were smokers. An equal number of controls were selected and of these, only 462 were smokers.

1.1 Write the 2x2 table.

(Exposure state)	(Outcome status)		Total
Cigarette smoking	Lung Cancer.		
	Present	Absent.	
Yes	499	462	961
No	19	56	75
Total	518	518.	

$q_1 = 100 - 51.93$
48.07 $q_2 = 100 - 25.33$
74.67

1.2 Test the hypothesis that smoking is associated with lung cancer.

Null hypothesis – There is no association between cigarette & lung cancer.

Percentage of exposed people from who got the disease $= \frac{499}{518} \times 100 = 96.33$

Percentage of ~~disease of~~ non exposed people who didn't get the disease $= \frac{462}{518} \times 100 = 89.19$

SND ≈ ?

$z = \frac{P_1 - P_2 - 0}{SE_{P_1 - P_2}}$

$SE_{(P_1 - P_2)} = \sqrt{\frac{P_1 q_1}{n_1} + \frac{P_2 q_2}{n_2}}$

$= \sqrt{\frac{96.33 \times 3.67}{518} + \frac{89.19 \times 10.81}{518}}$

$= 0.68 + 1.86 = \sqrt{2.54} = 1.59$

$z = \frac{96.33 - 89.19}{} = \frac{7.14}{1.59}$

$= 4.49$

95% confidence interval = 1.96. ∴ Null hypothesis.
Value is more than 1.96.

What you have done in 1.2 is testing a hypothesis for a case-control study using a **statistical** approach. Let us now consider testing this hypothesis using an **epidemiological** approach.

6.2 Odds Ratio

Recall how you tested a hypothesis in cohort studies by calculating a relative risk (RR) and attributable risk (AR) (pages 13-17).

1.3 Can you calculate RR and AR for the association in Exercise 1?

No. We can't.

The answer is NO.

You may recall that one needs to know the incidence rate to calculate the RR or IR. In a case-control study, where participants are selected on the basis of disease status, it is not possible to calculate the rate of development of a disease because the information on the population-at-risk (i.e. (a + b) or (c + d)) is not available.

However, the RR can be estimated by calculating the ratio of the odds of exposure among the cases to the odds of exposure among the controls. Let us see how this is derived.

If we consider a large population-at-risk, the number of people with a particular disease is likely to be very small compared to the number of people without that disease. Then, it follows that (refer to table 1):

(a + b) will closely approximate **b** and

(c + d) will closely approximate **d**

Accordingly, we can re-write the equation for RR as:

$$\text{Relative Risk} = \frac{a/b}{c/d} = \frac{a * d}{c * b}$$

We call this the **Odds Ratio (OR).**

$$\boxed{\textbf{Odds Ratio (OR)} = \frac{\mathbf{ad}}{\mathbf{bc}}}$$

1.4 Calculate the OR for the above exercise.

$OR = \frac{ad}{bc} = \frac{499 \times 56}{462 \times 19} = 3.18$

It should be clear to you now that in a case-control study, we can only calculate OR and not RR nor AR. However, the OR is interpreted in the same way as the relative risk.

> Odds Ratio (OR) is an approximate assessment of the relative risk.

1.5 Comment on the result.

There is 3.18 times as higher risk of smokers to get lung diseases ~~rather~~ compared than non-Smokers

Smokers are 3.18 times at more risk in getting lung diseases.

6.3 Advantages and disadvantages of case-control studies

Advantages:

- Ideal for identifying risk factors for rare diseases and also for diseases with a long incubation period.
- Relatively efficient requiring a smaller sample than a cohort.
- Relatively cheap.
- Can obtain results relatively quickly.
- Attrition (loss to follow up) is not a problem.
- Can investigate a wide range of possible risk factors.
- Consistency of measurement techniques can be easily maintained.
- Sometimes, this is the most feasible observational strategy for examining an association.

Disadvantages:

- Possible bias in selecting cases and controls (selection bias).
- Possible bias in measurement of exposure (recall and interviewer bias) and difficulties in obtaining the necessary information.
- There is no epidemiological denominator (population at risk) and therefore calculation of incidence rates is not possible.
- There is no way of finding out whether the exposure was the same for those who died and those who survived i.e. selective survival operates in case-control studies.
- It is not possible to find out about the pathology of other diseases related to the risk factor under investigation.
- In many case-control studies, there may be problems in sorting out the sequence of events (temporal sequence) i.e. whether the exposure led to the disease or vice versa.

7. Cross-sectional design in analytical studies

7.1 Design of a cross-sectional study

You may recall the cross-sectional surveys that we described under descriptive studies in volume 1 (pages 36-38). A cross-sectional design measures the prevalence of a disease (i.e. existing cases) and is useful in assessing the distribution of a disease and for formulating hypothesis on diseases and their exposures.

Cross-sectional design can also be used for examining associations in an analytical study that measures disease and exposure at the same point in time. Given below is the design for such a study.

Figure 1: Design of a cross-sectional study

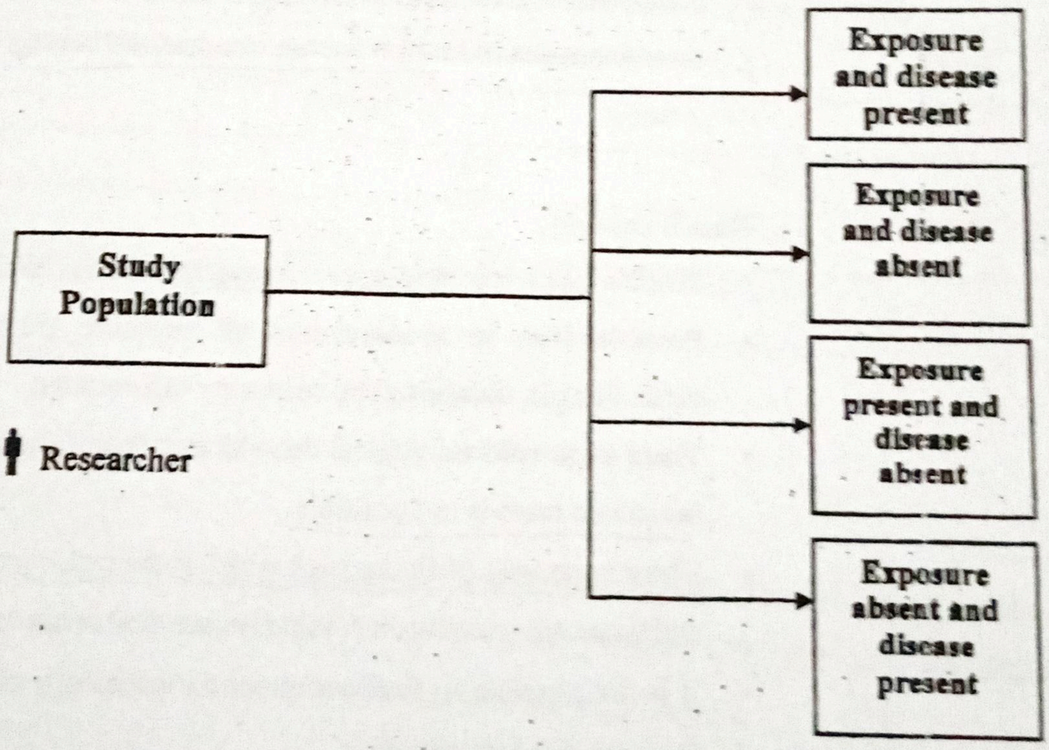


Figure 1 illustrates the design of a cross-sectional study. As shown, the individuals surveyed in this survey will fall into four categories. Comparison between the group with the disease and that without the disease with respect to the exposure variable will determine if a given exposure is associated with the disease under study.

Relationship between the exposure variable and disease can be presented in a 2x2 table, as shown in Table 1.

Table 1. Presentation of data observed in a cross-sectional study in a 2x2 table

Exposure	Outcome		Total
	Present	Absent	
Yes	a	b	a + b
No	c	d	c + d
Total	a + c	b + d	N

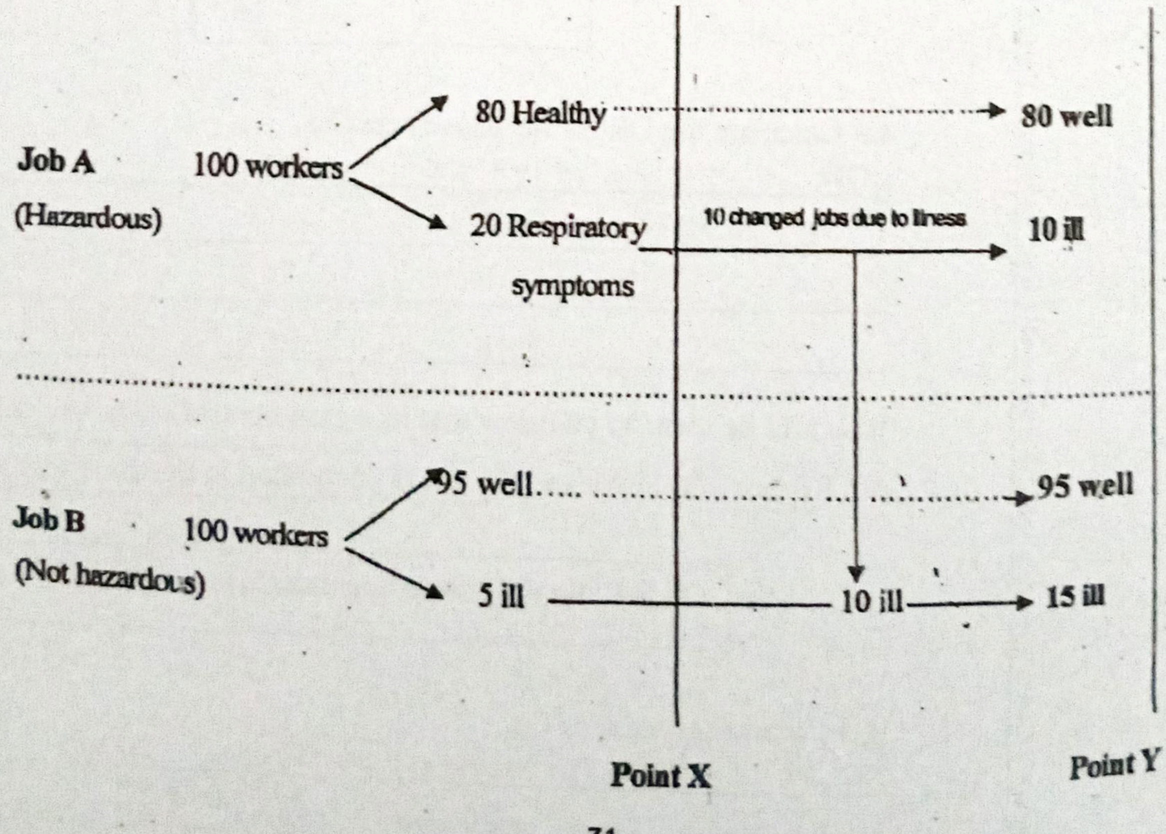
7.2 Advantages and disadvantages of cross-sectional analytical studies

A major strength of cross-sectional studies over cohort studies is that there is no waiting for the outcome to occur. Therefore, it makes them more feasible and less costly. In addition, it is the only study that gives the prevalence of a disease or risk factor.

However, there is one major weakness in this study design. Since both the exposure and disease status are measured at the same time, it makes it difficult to determine whether the exposure came before the disease or after the disease. This is clearly shown in the following example.

Example

Figure 1. Hypothetical illustration of the interrelationship between and occupational exposure and prevalence of disease as measured by a cross sectional study



1. Calculate the prevalence rates of respiratory symptoms in the two jobs at points X and Y.

Prevalence $A_x = \frac{20}{100} \times 100 = 20\%$

$B_x = \frac{5}{100} \times 100 = 5\%$

$A_y = \frac{10}{90} \times 100 = 11.1\%$

$B_y = \frac{15}{110} \times 100 = 13.6\%$

2. Calculate the ratio of prevalence rates between job A and job B at points X and Y.

~~Prevalence rate~~ = $\frac{20}{100}$ Point X prevalence ratio A:B = 20:5 = 4:1

Point Y Prevalence ratio = 11.1 ÷ 13.6 = 0.82 between A & B.

Based on the ratios of prevalence rates that you calculated at point Y, we may conclude that job B is more hazardous than job A. However, it is not so as it is because of the movement of affected workers from job A to job B. Therefore, being in job B would be the effect of those symptoms that they developed while in job A and not the hazardousness of job B.

This type of 'chicken or egg' dilemma is common when one uses a cross-sectional design in analytical studies.

Conclusions drawn on associations studied in cross-sectional studies need to be interpreted **with caution** taking into account the change in exposure following the outcome.

However, under special circumstances this study design can be used as effectively as in a prospective cohort study to test hypothesis on disease associations. That is, when the exposure variable does not change over time. Such variables include factors present at birth such as skin colour, eye colour, blood group and factors such as sex and highest level of education.

> Cross-sectional designs are more appropriate for measuring relationships between permanent characteristics of individuals and chronic diseases or stable conditions.

Cross sectional study designs are also impractical for the study of rare diseases, conditions of short duration and diseases with high case fatality, which are not detected by the one-time 'snap shot' of a cross-sectional study.

Exercise 1

In a cross-sectional survey of nutritional status among new school entrants, the heights of 853 children (558 males) were measured. It was found that 51 males and 41 of the female were stunted according to the Waterlow classification.

1.1 Write the 2x2 table.

	Stunting		
Gender	Present	absent	total
Male	51	507	558
Female	41	254	295
Total	92	761	853

1.2 Test the hypothesis that female students are at greater risk of stunting than males.

Proportion of stunted in males = $\frac{51}{558} \times 100\% = 9.1\%$

Proportion of stunted in females = $\frac{41}{295} \times 100\% = 13.8\%$

$SE = \frac{P_1(1-P_1)}{n_1} + \frac{P_2(1-P_2)}{n_2}$

P value = 0.13

odd's ratio = $\frac{41 \times 507}{51 \times 254}$ = 1.6.

Females have 1.6 times higher vulnerability for stunting

for 2 or mor proportions (T test or SND for 2 means)

7.3 Chi-Square Test

In Chapter 2, the data in Q1, Exercise 1 was presented in a table (these tables had 4 cells and was referred to as a 2x2 contingency table). In this exercise, we looked at the association of two events such as the exposure and outcome in **two independent samples.** We also used a statistical approach to find the significance of this association between the two events. This approach was by way of comparing **two proportions** using the **Z test or the SND test.**

When a comparison has to be made between two events but for **more than 2 independent samples,** the Z test or the SND test cannot be applied. Instead, another statistical test by the name of **chi-square test** (χ^2) has to be applied.

Exercise 2

Let us consider three independent groups or samples of patients all suffering from a particular disease undergoing three different treatments T_1, T_2 and T_3 to see how they respond to each treatment. If 100 patients given T_1, 70 (70%) responded favourably, 200 patients given T_2, 60 (30%) responded favorably and another 200 patients given T_3, 100 (50%) responded favourably, how do we find out the association between treatment and response to treatment?

In order to test this hypothesis on the association between treatment and response, we need to compare the proportions of patients responding favourably to these three treatments. In other words, a comparison has to be made between these three proportions. In this instance, we will apply the chi-square test to assess the significance of this association. Let us come back to this problem at a later stage in this chapter.

Let us first apply the chi-square statistic for a two sample situation, as given below.

Exercise 3

To find the association between gestational diabetes mellitus (GDM) and the parity of women, an obstetrical study was conducted among 790 expected mothers of 30 years of age. Of these, 480 were in their first pregnancy with 30 of them having GDM, while only 12 of the remaining pregnant women having GDM. Is there any association between gestational diabetes and the parity of women?

To apply the chi-square test to this above situation, the following steps have to be followed.

i. Complete the table given below using the information in Exercise 3. These values are called **'observed frequencies'**.

Table 2. Observed values for the association between GDM and parity of women

Exposure	**Outcome**		**Total**
	GDM +	GDM -	
First pregnancy	30 a	450 b	480 a+b
Subsequent pregnancies	12 c	298 d	310 c+d
Total	42 a+c	748 b+d	790 N

ii. The chi-square test is based on calculating a set of **expected values** for each cell in the above table. The expected values are based on the assumption that the null hypothesis of 'There is no association between the study groups in relation to the proportion of the factor of interest' is true. If there was no association between the exposure (parity) and disease (GDM), the proportion of cases that were exposed would be the same as the proportion of the entire study population that is exposed.

If we want to calculate the expected number of primi-parous women with GDM (a), it would be = [(a+b) / N] * (a+c)

☞ Refer Table 1 in Chapter 2 (page 10) for these symbols.